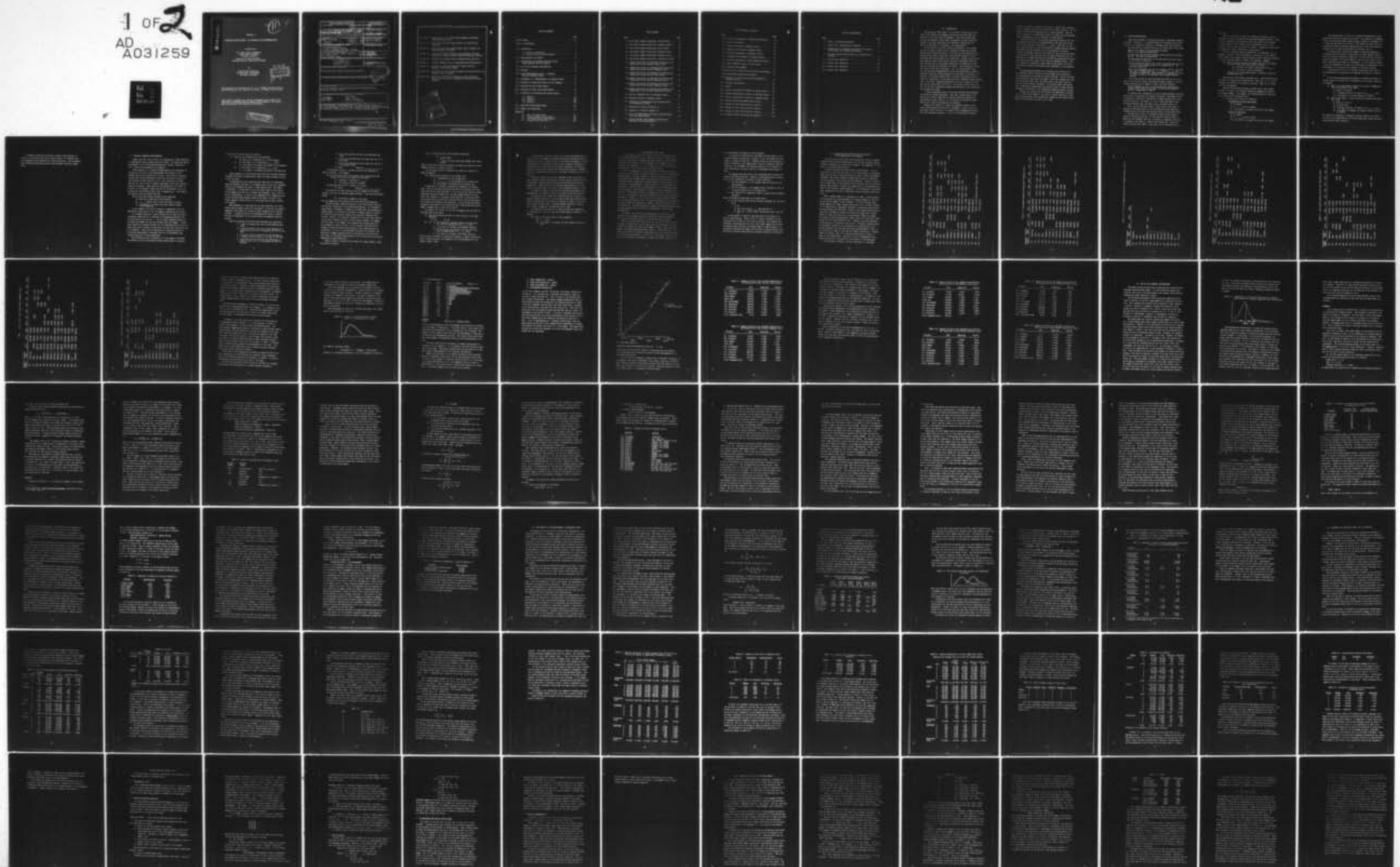
AD-A031 259 MICHIGAN UNIV ANN ARBOR COASTAL ZONE LAB F/G 13/2
PILOT STUDY PROGRAM, GREAT LAKES SHORELAND DAMAGE STUDY. APPEND--ETC(U)
MAY 76 E D ROTHMAN, S BARTOLD, R BUTLER DACW23-75-C-0027

UNCLASSIFIED NL

1 OF 2
AD
A031259



(11)

APPENDIX V

SHORELINE DAMAGE SURVEY: AN APPRAISAL WITH RECOMMENDATIONS

Prepared for

U.S. Army Corps of Engineers
North Central Division
Chicago, Illinois

Contract No. DAC23-75-C-0027
Reference Contract Modification P00001

By

Coastal Zone Laboratory
The University of Michigan
Ann Arbor, Michigan

Prepared under the supervision of Dr. E. D. Rothman, Associate Professor of Statistics, by S. Bartold, R. Butler, M. Coffey, and L-W Huang.

This report is submitted in partial fulfillment of the terms of The University of Michigan Contract Number 011456 awarded to the Department of Statistics, The University of Michigan.

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER

2. GOVT ACCESSION NO. 6

3. RECIPIENT'S CATALOG NUMBER 9

4. TITLE (and Subtitle) ~~Summary Report of the~~ Pilot Study Program, Great Lakes Shoreland Damage Study. [redacted] Appendix V. Shoreline Damage Survey: An Appraisal with Recommendations.

5. TYPE OF REPORT & PERIOD COVERED
Final rept.

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)
10
[redacted] E. D. Rothman, M. Coffey
S. Bartold, L-W Huang
R. Butler

8. CONTRACT OR GRANT NUMBER(s)
15 DACW 23-75-C-0027

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Coastal Zone Laboratory
1101 North University Building
University of Michigan, Ann Arbor, MI 48104

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

11. CONTROLLING OFFICE NAME AND ADDRESS
North Central Division, Corps of Engineers
536 S. Clark Street
Chicago, Illinois 60605

12. REPORT DATE
11 May [redacted]976

13. NUMBER OF PAGES
145

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)
12 149p.

15. SECURITY CLASS. (of this report)
Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)
Approved for public release, distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES
Copies are obtainable from National Technical Information Service, Springfield, Virginia 22151

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

flood damage
erosion damage
coastal zone
Great Lakes
data collection
statistical research

This appendix

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)
This is an appendix to the Summary Report of the Pilot Study Program, Great Lakes Shoreland Damage Study. It is an appraisal of the data collection techniques and statistical procedures used in the six Michigan counties studied.

409 887 LB

Main Report	Summary Report of the Pilot Study Program, Great Lakes Shoreland Damage Study.
Appendix I	Great Lakes Shoreline Damage Survey; St. Louis County, Minnesota
Appendix II	Great Lakes Shoreline Damage Survey; Brown, Douglas, and Racine Counties, Wisconsin.
Appendix III	Great Lakes Shoreline Damage Survey; Muskegon, Manistee, Schoolcraft, Chippewa, Alcona, and Huron Counties, Michigan
Appendix IV	Contract for a Damage Survey of Oswego County, New York
Appendix V	Shoreline Damage Survey: An Appraisal with Recommendations
Appendix VI	Engineering - Economic Analysis of Shore Protection Systems: A Benefit/Cost Model
Appendix VII	Measurement of Coastal Bluff Recession from Aerial Photographs, Muskegon County, Michigan
Appendix VIII	Comparison of Field Data Collection to Date Collected Using Study Instruments in Muskegon and Manistee Counties, Michigan

TABLE OF CONTENTS

LIST OF TABLES

LIST OF TABLES, continued

LIST OF ILLUSTRATIONS

1.0 INTRODUCTION

The purpose of this report is to provide a constructive criticism of the coastal zone damage survey being conducted under the auspices of the Army Corps of Engineers. With this general goal in mind, we examine in this report the methods of data collection, coding of this data and discuss the methods of summarizing the data in a form which is convenient, yet comprehensive.

We turn first to the data collection procedure. The primary aim was to obtain a census of all residents of shoreline property within each of several counties. A prelist of these properties was obtained from county records and a questionnaire mailed to every household (with Huron County as an exception). Though certain errors were found to exist on the prelist (several respondents indicated that they were not in the target population) the magnitude of this error is not known but believed to be small. More substantive errors were found in the basic instrument, i.e., the questionnaire form. Our criticism and recommendations on how the questionnaire should be modified along with coding procedures are discussed in Section 2. In addition to the mailed questionnaire, personal interviews of both a sample of respondents and non-respondents were conducted using a second "interview" form. Suggested improvements for the collection of this data are also discussed in Section 2. Criticism, in this case, centers on the lack of agreement between the questions of the "interview" and questionnaire forms.

A final independent source of data was obtained from assessment rolls and real estate appraisals. This data, though useful in assessing the relative magnitudes of numbers generated through the questionnaire format, was not, in our opinion, sufficiently comprehensive as a basis for substantive conclusions.

Though we are aware of a number of arguments (based on both social and political concerns) for attempting a census as the primary mode of data collection, we believe that a sampling plan would be a more efficient procedure. This plan is described in Section

9 and is, in brief, a census in the case of "small" rural counties while a sampling procedure is advocated for larger counties. Our argument for using this sampling plan is simply economic. That is, more information, of comparable quality (as judged from the decision makers perspective), may be obtained for fewer dollars.

Both a criticism of the proposed descriptive measures for summarizing the data and our recommendations on how this may be done effectively are given in Sections 3, 4, and 5. The first two sections deal, indirectly, with the question of summarizing the observed data while the final section treats the question of outliers. Clearly, of primary concern is the quality of the raw data and hence the ordering of these sections does not imply differing levels of importance. A comparison between interview and questionnaire results is given in Section 6 based on the data for Muskegon County. This is followed by a comparison, again based on data drawn from this same county, between nonrespondents and respondents to the questionnaire.

The methodology necessary to extrapolate the evaluation of the expanded totals to the entire county based on the results of the questionnaire is provided in several sections. Section 8 treats the problem of projecting total damage, total cost (damage plus cost of protection), bluff lost, and beach lost at the reach level for each of the six counties surveyed. On the other hand, Sections 10 and 11 discuss the methodology that would be used to handle similar problems of prediction if the recommended sampling proposal is used.

Finally, our results and conclusions are summarized in Section 13. In view of the length of this report, direct reference to this section may provide a more appropriate roadmap for the reader interested in only specific details of our study.

1.1 MAILED QUESTIONNAIRE

There were problems of interpretation when a respondent answered both A2a, b and A2c, d. A possible means for clearing up such difficulties is to change the questions to something like the following:

How many distinct properties do you own, which front on the lake?

What is the length of your shoreline frontage?
(Total or by separate properties??)

How many feet back from the present shoreline does your property extend? (By separate properties??)

How many distinct properties in the lake area do you own, which do not front on the lake?

For your properties which do not front on the lake, how far from the lake is your property line at the nearest point? (by separate properties?)

What is the approximate size of your property; that is, how many feet does it measure each way? ______ feet by ______ feet. (By separate properties? Indicate which is frontage and which is depth, to assure correspondence with frontage question above??)

If you only include in the study properties which front on the lake, you may wish to ensure that property owners respond to the following questions only in regard to their lake front property:

Please complete the rest of the questionnaire regarding only property which fronts on the lake. If you own no lakefront property, please skip to page 8.

In question A6a, information is lost by coding different dwelling structures on the property as separate variables. Information would be retained by coding the responses as a single variable with classifications such as: one house, two houses, house and mobile home, cottage and mobile home, one mobile home, two mobile homes, etc.

Also, for all questions, make it clear that if more space is needed, an additional sheet may be enclosed giving the additional information (such as additional dwelling units, if more than three).

The same sort of loss of information exists in the present method of coding A6b and A6c. Perhaps A6a-A6c could be coded together as one variable, such as: one house, 1, seasonal; one house, 1, permanent; one house, 1, seasonal and income; etc., where the number indicates the number of dwelling units in the structure it

follows.

What we would like is that all categorical variables such as A6a-A6c be coded in a mutually exclusive and exhaustive manner. That is, any individual case belongs to only one level of the variable and the levels of the variable cover all possible cases. This should be done in such a way that no useful information is lost (such as number and type of dwellings on each property).

For all questions, make it difficult for a respondent not to answer a question. Prior formulation of the question will make it much easier to distinguish between missing values, those with answers of zero, and those unable to answer the question. Example:

A8. What would you estimate the market value of your property to be if normal lake levels were to return?

☐ I estimate the market value would be $_______.

☐ I am unable to make an estimate.

(Upon what is the estimate based? If unable to make an estimate, why not?)

Approximately forty people answered yes to question A9 but did not bother to answer any of question B2 or the cost portions of B4 (damage costs and cost of protective actions). What is the reason for this discrepancy?

Question B2 involves estimates of flood and erosion damage. We need to be able to distinguish between no damage and inability to estimate damage. For example, the different parts of the question could have the following form:

a. Structure and contents of residence

Flooding due to high lake levels

☐ No damage.

☐ $________ worth of damage.

☐ I am unable to estimate the cost of the damage.

Erosion of shoreline

☐ No damage.

☐ $ ________ worth of damage

☐ I am unable to estimate the cost of the damage.

For questions B4a, b, and c try to make the questions so that people who answer one answer all of these. For instance, in Muskegon County, why did seven people answer Action 3, while 21 estimated cost of labor and materials for Action 3? This seems odd since there is consistency between those who answered Actions 1 and 2 and those who answered cost of labor and materials for Actions 1 and 2.

For question B5, increase the number of possible responses to correspond with the present coding format, so that people's answers fall into mutually exclusive categories. If these mutually exclusive categories appear on the questionnaire, then we do not need to worry when interpreting the number of respondents in each category whether or not the respondents all knew they could check all appropriate categories(which is the case as the question is currently stated).

In questions C7, C8, and D5 include "since Labor Day, 1972" in the formulation of the question (or whatever the period of interest is).

In the questions dealing with danger of flooding and danger of erosion, make it difficult for a respondent not to answer a question. For example:

D1. What is the approximate height of the bluff or embankment above the existing water level?

- ☐ There is no bluff.
- ☐ I estimate distance "A" above to be ________ feet.
- ☐ I am unable to estimate distance "A" above.

D3. What is the depth of bluff loss due to erosion since Labor Day, 1972?

- ☐ There is no bluff.
- ☐ No bluff loss.
- ☐ I estiamte distance "C" above to be ________ feet.
- ☐ I am unable to estimate the depth of bluff loss due to erosion.

Our reason for including the possible response "there is no bluff," is to allow for people who do not believe they have a bluff and might otherwise skip these questions.

In regard to questions B4 and E1, exclude flood insurance as a type of protective action that could be taken. An enlarged question regarding flood insurance was listed previously. See recommendations regarding protective actions under personal interview suggestions.

1.2 PERSONAL INTERVIEW QUESTIONNAIRE

Again, with the current form, it is impossible on many questions to distinguish between missing and zero values. The questions should be phrased so as to distinguish between zero values, inability to give a figure and skipping of the question.

After question 8, it should be asked whether the respondent is filling out the rest of the questionnaire with respect to all housing units on the property or just one. With regard to question 10 and other similar questions, if you are interested not only in the building in which they live (i.e., also interested in those they own but rent or are at present unoccupied), then this question should be restated. Also, allow for the possibility that there is no house on the property (since you ask the interviewer to put one housing unit per interview form, this is appropriate). For example, the question could be worded this way:

What is the total square footage of this dwelling?

- ☐ There is no dwelling on this property.
- ☐ ________ square feet is the total square footage.
- ☐ I am unable to estimate the total square footage of this dwelling.

Is the comment "if you don't know go to questions 11 and 12" appropriate after question 10? We suggest eliminating this comment

If a separate interview form is filled out for each dwelling on a property, make sure that responses not dealing specifically with the dwelling are not duplicated. All of these responses should have the same case number so that this mistake could not happen. Perhaps there should be a special variable for the dwellings (with no pre-set upper limit such as three, in case there are more than three dwellings on a property). Also, see the discussion of questions A6a-A6c on the mailed questionnaire concerning the loss of information when related variables are coded separately.

The following version of question 11 is an example of wording the question so as to distinguish between those unable to estimate

a value and those who skip the question.

11. What are the dimensions of this dwelling?

☐ There is no dwelling unit on this property.

a) ☐ The length of the dwelling is _____ feet.

☐ I am unable to estimate the length of the dwelling.

b) ☐ The width of the dwelling is _____ feet.

☐ I am unable to estimate the width of the dwelling.

After question 12, ask whether each floor has the same dimensions. Variable 407 gives an inappropriate value for square footage if this is not the case.

In question 15, does "rented by you" mean that the respondent owns the property and rents it to someone else or that the respondent lives on the property and rents it from someone else? This makes a difference. A tenant may be more (or less?) aware of the extent of the damage and loss due to erosion and flooding. Also, later on in the questionnaire, it may be that expenditures for damage and protective actions were primarily made by the owner, so that if the renter answers a biased view of the situation may be given (biased up or down?).

What happened if, when going out for the interview, the interviewer found a tenant rather than the owner? If you get the tenant's responses, this should be carefully noted, possibly with a special variable. A question such as the following would help to classify the type of respondent (if you allow tenants to complete questionnaires):

Check the response below which best describes your situation:

☐ I own the property and live in this dwelling year round.

☐ I own the property and live in this dwelling on a seasonal basis and it is unoccupied the rest of the year.

☐ I own this property and live in this dwelling on a seasonal basis and rent it the rest of the year.

☐ I own this property and rent this dwelling on a seasonal basis and it is unoccupied the rest of the year.

☐ I own this property and rent this dwelling year round.

☐ I rent this dwelling from the owner and live in it year round.

☐ I rent this dwelling from the owner and live in it on a seasonal basis.

☐ Other Describe ______________.

In questions 16 and 17, include an option for inability to make an estimate. For example:

16) If your were to sell your property now, during high lake levels, how much do you think you could get?

☐ I estimate I would get $ _______.

☐ I am unable to make an estimate.

In questions 16a and 17a, include as a choice:

☐ Other Describe ______________.

For question 19, add clean-up costs and damage to the septic system to the list of possible damage suffered. Also add the possible answer:

☐ I suffered no damage of this type.

See the suggestions for question B2 of the mailed questionnaire.

We had problems comparing question 19 and those which follow it with similar questions on the mailed questionnaire. On the mailed questionnaire, estimates of damage were broken down into estimates of damage caused by flooding and estimates of damage caused by erosion. The interview questionnaire seems to be primarily concerned with erosion. However, the questions (especially 19) do not state specifically that estimates are only to cover erosion damage. We suggest that the cause of the damage to be estimated be stated explicitly as flooding or erosion or both. If the personal interview form is changed to match the mailed questionnaire, this problem should be eliminated. As it is, it was extremely difficult to interpret comparisons of damage estimates given on the personal interview form with those on the mailed questionnaire, since we were not sure we were comparing the same quantities.

Again, include in question 20 an option for those unable to give the desired response:

20) On what dates were these damages experienced?

☐ ______________
month year

☐ I cannot estimate when these damages were experienced.

Make it clear that multiple responses for dates are appropriate when damage occurred on several occasions.

The following version of question 21 allows for inability to make an estimate:

21) What is your estimate of net income lost?

☐ I experienced no net income loss.

☐ $____________ is my estimate of net income lost.

☐ I am unable to estimate the net income lost.

In question 22, include a response for inability to estimate cost. Also, do you want protective actions for any dates, not just Labor Day 1972 to Labor Day 1974? If so, how can the costs for these protective actions be added onto the above damage figures to give total cost figures for the period Labor Day 1972 to Labor Day 1974? If you want costs for protective actions only for the stated period, this should be included in the questions. If you do want protective actions for any dates, do you want the cost of any protective action added to damage costs to give total costs?

In question 22a, after (1) and (2), we suggest that you add the following additional parts:

(3) Was the building relocated to a new location on the same property?

☐ Yes ☐ No

(4) (i) ☐ The building's original distance from the beach (or bluff as appropriate) was ____ feet.

☐ I am unable to estimate the original distance.

☐ The building's new distance from the beach (or bluff as appropriate) is ________ feet.

☐ I am unable to estimate the current distance.

Part (4) may be a means of assessing the degree of danger from flooding or erosion--how far the respondent felt the building had to be moved to make it safe.

In questions 22 and 23, as well as in B4 of the mailed questionnaire, a list of examples of possible protective actions would be appropriate, such as: armor the toe of the bluff, entrainment of shoreline materials, dissipation of wave energy offshore, replacement of beach materials, relocation of buildings, evacuation of buildings, modify the flood plain, modify the flood plain structure. Examples help to jog memories and illustrate the types of efforts you are looking for.

Questions 22 and 23 should be combined so that a cost and success response could be associated with each type of protective action, rather than a lump sum given and then actions and success listed separately. (Similar to B4 on the mailed questionnaire, but with room for as many types of actions as the respondent wishes to list).

In regard to question 23B, why not have the interviewer take a photograph of the property owner's shore protection structure(s)?

For question 23E, rather than allowing only for a yes/no response, include boxes for a listing such as that which follows which gives an indication of the degree of success (or lack of it) resulting from the actions taken: permanent/good to excellent, limited/ fair to poor, temporary, none/adverse, don't know.

For questions 24-35 and 37-44, again provide the opportunity for respondents to indicate that they cannot make an estimate. For example:

24) What is the total depth of this property?

☐ ______ feet

☐ I am unable to estimate the total depth of this property.

2.0 QUESTIONNAIRES AND CODING

In the pages which follow in this section, we suggest ways in which the questionnaires can be improved in order to obtain more accurate assessments of damages and losses and to prevent misinterpretations of questions by the respondent and responses by the analyst. See Appendix V-c for a copy of the mailed questionnaire and Appendix V-d for the personal interview form.

The first suggestion is that the questions which are asked on the mailed questionnaire be asked in identical form on the interview questionnaire. The main reason for conducting personal interviews with persons who already filled out the mailed questionnaire is to determine whether people give the same answers to the written as to the interview questionnaire. Unless questions to be compared are identical in both settings, interpretations of any comparisons made must be suspect. For instance, we wish to compare respondents' estimates of total damage in written and interview form. Does the presence of an interviewer cause the respondent to give more conservative estimates of losses, or vice versa, or are answers similar in both written and interview settings. A proper assessment of such questions can only be made if questions are asked in identical form on both questionnaires. The same discussion applies when comparing interviews of nonrespondents with interviews of respondents to the mailed questionnaire.

Of course, the interview questionnaire may contain more questions if this is deemed appropriate, but they should be separate from the questions which appear in the mailed questionnaire.

During the interviews which occur on the site of the property, it might be appropriate for the interviewer and respondent to walk around the property and assess the damage together. Also, the interviewer could make some measurements himself. This would serve as a check on accuracy of responses (measurements on site versus responses on the mailed questionnaire) and give the interviewer a feel for how well the respondent's assessments of damage and losses compare with

the interviewer's perception of the situation.

In analysis of questionnaire responses, we were unable to distinguish between missing data, answers of zero, and people who were unable to answer the question. Examples follow in the discussion of specific questions, with suggestions of ways to eliminate these difficulties. We recommend that missing values for all variables be coded identically, in such a way as to distinguish from answers of zero.

We suggest that some form of the following three questions be added to the questionnaires. Answers to them should give an indication of the reliability of answers to other questions.

1) Was the respondent at the property when he filled out the questionnaire?
2) Did the respondent claim damage (flood, erosion) or cost of protective measures as a tax deduction?
3) How long has the respondent owned or occupied this lakeshore property?

Also, the question below might be of some value:

4) Do you currently have flood insurance coverage for this property?
   - ☐ No
   - ☐ Yes, and I pay $______ per year for it.
   - ☐ Yes, but I cannot recall how much per year I pay for it.

The discussion of specific questions below is divided up by type of questionnaire. This is only for ease of reference to the questionnaires as used for the present analyses. We recommend that subsequent questionnaires be written with these suggestions in mind, eliminating some questions and adding others to improve quality of information obtained and remove difficulties in interpretation, with mailed and interview questionnaires as nearly identical as possible.

3.0 DISTRIBUTIONS OF VARIABLES FROM THE MAILED QUESTIONNAIRE FOR SIX COUNTIES

In this section we describe the distributions of the variables obtained from the mailed questionnaire for Alcona, Chippewa, Huron, Manistee, Muskegon, and Schoolcraft Counties. See Appendix V-c for a copy of the mailed questionnaire.

Fourteen variables have multinomial distributions. That is, responses fit into one and only one of a finite number of categories. There are problems of interpretation with some of these multinomial variables. For these questions, people were allowed to make more than one response, as in B5 of the questionnaire. However, since it was not made clear in the wording of the questions that more than one response was allowed, it cannot be assumed that all respondents were aware of the intent of these questions. See the section on questionnaire suggestions for ways of improving these multinomial type questions.

When a random sample of a population is taken, a good estimate of the proportion of the population in a given category is the proportion of the sample in that category. In future surveys, when random samples are taken, sample proportions such as those in Tables 3.1-3.5 may serve as reasonable estimates of population proportions for multinomial type variables. However, such sample proportions are reasonable estimates of population proportions only when a random sample is taken and there are no nonrespondents among the sample. The reason is that nonrespondents (people who do not return the questionnaire) are often different from respondents for many important characteristics of interest in the survey. Therefore, answers given by respondents should not be considered as representative of an entire population if there were very many nonrespondents. This is why care must be taken to see that all people selected for a random sample do respond, even if it requires telephoning or visits to homes of the recalcitrant. See the section on recommended future sampling plans for a discussion of the problem of nonrespondents.

In the present situation, the 305 respondents for Muskegon

TABLE 3.1 Multinomial Sample Proportions, Alcona County

Variable Number	Variable Name	$\hat{p}(-1)$	$\hat{p}(0)$	$\hat{p}(1)$	$\hat{p}(2)$	$\hat{p}(3)$	$\hat{p}(4)$	$\hat{p}(5)$	$\hat{p}(6)$	$\hat{p}(7)$	$\hat{p}(8)$	$\hat{p}(14)$
V8	Dwelling	8/552		404/552	140/552							
V9	A6A1	148/552		199/552	189/552	3/552	10/552		1/552		2/552	
V10	A6A2	493/552		8/552	28/552	1/552	2/552		18/552		2/552	
V11	A6A3	537/552		2/552	8/552				1/552	1/552	3/552	
V12	Nodwell 1		152/552	392/552	4/552			2/552	1/552			1/552
V13	Nodwell 2		511/552	35/552	1/552		2/552	1/552	2/552			
V14	Nodwell 3		541/552	7/552	2/552		1/552	1/552				
V15	Dweltyp 1	146/552		267/552	125/552	4/552	5/552	4/552	1/552			
V16	Dweltyp 2	502/552		26/552	11/552	10/552	2/552	1/552				
V17	Dweltyp 3	540/552		2/552	3/552	6/552	1/552					
V20	Protect?		8/552	288/552	256/552							
V22	Damage?	155/552		199/552	198/552							
V60	B5	163/552		79/552	27/552	253/552	3/552	11/552	16/552			
V75	Floodins	32/552		19/552	501/552							

TABLE 3.2 Multinomial Sampling Proportions, Chippewa County

Variable Number	Variable Name	$\hat{p}(-1)$	$\hat{p}(0)$	$\hat{p}(1)$	$\hat{p}(2)$	$\hat{p}(3)$	$\hat{p}(4)$	$\hat{p}(5)$	$\hat{p}(6)$	$\hat{p}(7)$	$\hat{p}(8)$
V8	Dwelling	9/605		445/605	151/605						
V9	A6A1	162/605		145/605	249/605	8/605	32/605	1/605	2/605	1/605	5/605
V10	A6A2	489/605		4/605	47/605	5/605	9/605	1/605	23/605	4/605	23/605
V11	A6A3	567/605		2/605	7/605	2/605	3/605		18/605	2/605	4/605
V12	Nodwell 1		180/605	418/605	3/605		1/605		1/605		
V13	Nodwell 2		531/605	61/605	3/605	2/605	4/605	1/605			2/605
V14	Nodwell 3		589/605	12/605	3/605	1/605					
V15	Dweltyp 1	164/605		324/605	97/605	10/605	6/605	4/605			
V16	Dweltyp 2	520/605		58/605	9/605	16/605	2/605				
V17	Dweltyp 3	582/605		12/605	4/605	5/605	2/605				
V20	Protect?	24/605		449/605	132/605						
V22	Damage?	89/605		414/605	102/605						
V60	B5	101/605		70/605	23/605	349/605		11/605	51/605		
V75	Floodins	30/605		10/605	565/605						

TABLE 3.2 Multinomial Sampling Proportions, Chippewa County (cont'd)

Variable Number	Variable Name	$\hat{p}(10)$	$\hat{p}(11)$	$\hat{p}(20)$
V8	Dwelling			
V9	A6A1			
V10	A6A2			
V11	A6A3			
V12	Nodwell 1	1/605	1/605	
V13	Nodwell 2			1/605
V14	Nodwell 3			
V15	Dweltyp 1			
V16	Dweltyp 2			
V17	Dweltyp 3			
V20	Protect?			
V22	Damage?			
V60	B5			
V75	Floodins			

TABLE 3.3 Multinomial Sample Proportions, Huron County

Variable Number	Variable Name	$\hat{p}(-1)$	$\hat{p}(0)$	$\hat{p}(1)$	$\hat{p}(2)$	$\hat{p}(3)$	$\hat{p}(4)$	$\hat{p}(5)$	$\hat{p}(6)$	$\hat{p}(7)$	$\hat{p}(8)$
V8	Dwelling	4/280		231/280	45/280						
V9	A6A1	50/280		121/280	101/280		8/280				
V10	A6A2	234/280		4/280	11/280	1/280	5/280		17/280		8/280
V11	A6A3	268/280		1/280	3/280				5/280		3/280
V12	Nodwell 1		67/280	210/280	3/280						
V13	Nodwell 2		260/280	18/280	1/280					1/280	
V14	Nodwell 3		276/280	4/280							
V15	Dweltyp 1	53/280		160/280	60/280	5/280	1	1			
V16	Dweltyp 2	257/280		16/280	2/280	5/280					
V17	Dweltyp 3	277/280		1/280	1/280	1/280					
V20	Protect?	8/280		185/280	87/280						
V22	Damage?	59/280		163/280	58/280						
V60	B5	68/280		21/280	10/280	161/280	1/280	3/280	16/280		
V75	Floodins	10/280		8/280	262/280						

TABLE 3.4 Multinomial Sample Proportions, Manistee County

Variable Number	Variable Name	$\hat{p}(-1)$	$\hat{p}(0)$	$\hat{p}(1)$	$\hat{p}(2)$	$\hat{p}(3)$	$\hat{p}(4)$	$\hat{p}(5)$	$\hat{p}(6)$	$\hat{p}(7)$	$\hat{p}(8)$	$\hat{p}(9)$
V8	Dwelling			132/204	72/204							
V9	A6A1	71/204		79/204	53/204					1/204		
V10	A6A2	189/204		4/204	5/204				3/204	2/204	1/204	
V11	A6A3	199/204		1/204	1/204				3/204			
V12	Nodwell 1		74/204	126/204	2/204	1/204						1/204
V13	Nodwell 2		190/204	12/204	1/204			1/204				
V14	Nodwell 3		201/204	2/204	1/204							
V15	Dweltyp 1	71/204		74/204	2/204	5/204	5/204	2/204				
V16	Dweltyp 2	191/204		8/204	2/204	2/204		1/204				
V17	Dweltyp 3	201/204		1/204	1/204			1/204				
V20	Protect?	2/204		175/204	27/204							
V22	Damage?	22/204		155/204	27/204							
V60	B5	20/204		40/204	6/204	132/204		2/204	4/204			
V75	Floodins	8/204		17/204	179/204							

TABLE 3.5 Multinomial Sample Proportions, Muskegon County

Variable Number	Variable Name	$\hat{p}(-1)$	$\hat{p}(1)$	$\hat{p}(2)$	$\hat{p}(3)$	$\hat{p}(4)$	$\hat{p}(5)$	$\hat{p}(6)$	$\hat{p}(7)$	$\hat{p}(8)$	$\hat{p}(9)$	$\hat{p}(50)$
V8	Dwelling		232/305	73/305								
V9	A6A1	79/305	128/305	95/305		1/305			1/305	1/305		
V10	A6A2	263/305	11/305	23/305				4/305	1/305	3/305		
V11	A6A3	299/305	1/305	2/305				2/305		1/305		
V12	Nodwell 1	83/305	216/305	1/305	3/305	1/305		1/305				
V13	Nodwell 2	270/305	29/305	1/305	1/305		1/305		1/305		1/305	1/305
V14	Nodwell 3	301/305	3/305		1/305							
V15	Dweltyp 1	76/305	105/305	108/305	8/305	6/305	2/305					
V16	Dweltyp 2	268/305	13/305	6/305	12/305	2/305	4/305					
V17	Dweltyp 3	300/305	1/305	2/305	1/305	1/305						
V20	Protect?	2/305	249/305	54/305								
V22	Damage?	9/305	239/305	57/305								
V60	B5	29/305	32/305	1/305	228/305		6/305	9/305				
V75	Floodins	11/305	45/305	249/305								

TABLE 3.6 Multinomial Sample Proportions, Schoolcraft County

Variable Number	Variable Name	$\hat{p}(-1)$	$\hat{p}(0)$	$\hat{p}(1)$	$\hat{p}(2)$	$\hat{p}(3)$	$\hat{p}(4)$	$\hat{p}(6)$	$\hat{p}(8)$	$\hat{p}(9)$	$\hat{p}(16)$
V8	Dwelling			58/134	76/134						
V9	A6A1	77/134		28/134	25/134		2/134		77/134		
V10	A6A2	126/134		1/134	4/134			2/134	1/134		
V11	A6A3	132/134			1/134				1/134		
V12	Nodwell 1		80/134	53/134							1/134
V13	Nodwell 2		128/134	6/134							
V14	Nodwell 3		132/134	2/134							
V15	Dweltyp 1	78/134		34/134	20/134	2/134					
V16	Dweltyp 2	128/134		4/134	1/134	1/134					
V17	Dweltyp 3	132/134		2/134							
V20	Protect?	5/134		32/134	97/134						
V22	Damage?	50/134		28/134	56/134						
V60	B5	76/134		34/134	3/134	26/134					
V75	Floodins	12/134		5/134	117/134						

County, for instance, represent approximately 60% of the lakeshore property owners in Muskegon County. All other property owners must be classified as nonrespondents since questionnaires were mailed to all property owners. Therefore, results concerning the respondents should be considered representative of all lakeshore property owners only if it is assumed that respondents and nonrespondents in Muskegon County are basically alike in the characteristics of interest. As stated above, such an assumption may lead to erroneous conclusions. Similar remarks apply to the other four counties being considered here.

The multinomial variables are listed in Tables 3.1-3.6 with the proportion of respondents within each category. If a category is not listed, then the proportion is 0. Note that the proportions listed are not proportions for the entire county, but rather proportions from the sample of respondents with which we were provided.

In Tables 3.1-3.6 the notation p(i) stands for the proportion of respondents within category i. For each variable, we have included a new category, -1, which covers all missing values. The questionnaire suggestions include a discussion of these missing values. Some missing values are from people who did not respond to the question, some from people who could not answer the question, and some from people for whom the question did not apply (for instance, property owners not having three dwellings on their lakeshore property would not be included in the count for variable 11, A6a3). Different types of missing values cannot be distinguished at this time. This problem can be alleviated with careful wording of future questionnaires and careful coding of responses.

There were so few responses for flooding variables 61-69 for the six counties that no distributions were found for these variables. Also, no distribution was found for variables 7, Reachno. No useful distribution could be found for variable 5, Frontage.

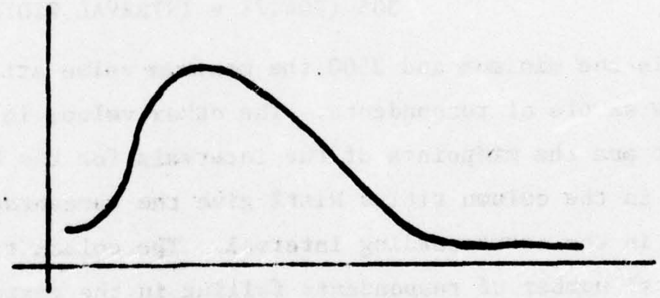
No useful distribution could be found for variable 19, Propdelt. This question (A8) was subject to so many possible interpreta-

tions that is was ignored in our analyses. See suggestions for improvement of this question under questionnaire suggestions.

The other variables to be considered are reasonably well described by the lognormal distribution. These variables are Assessed Value (variable 2 times variable 4, V2*V4); variable 6, Propdepth; variable 18, Propworth; Total Damage (the sum of variables 23-36, 38-43); Total Cost (Total Damage plus the cost of protection variables 46, 47, 51, 52, 57, 58); variable 70, Bluffheight; variable 71, Beachdepth; variable 72, Blufflost; variable 73, Bluffdist; and variable 74, Beachlost.

When histograms for these ten variables were made, each resembled a curve such as the following:

Figure 3.1 Graph of a Skewed Distribution With a Heavy Tail (Outliers) to the Right



For example, the Midas command

HISTOGRAM VAR=6 INTERVAL=* OPTION=HIST%

produced the following histogram as output for Muskegon County data.

HISTOGRAM/FREQUENCIES

MIDPOINT	HIST%	COUNT FOR PROPDEPT	(EACH X = 2)
20.000	6.1	18	XXXXXXXXX
224.71	22.9	68	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
429.41	13.8	41	XXXXXXXXXXXXXXXXXXXXX
634.12	17.5	52	XXXXXXXXXXXXXXXXXXXXXXXXXX
838.82	15.8	47	XXXXXXXXXXXXXXXXXXXXXXXX
1043.5	5.7	17	XXXXXXXXX
1248.2	4.7	14	XXXXXXX
1452.9	4.7	14	XXXXXXX
1657.6	2.4	7	XXXX
1862.4	1.7	5	XXX
2067.1	2.7	8	XXXX
2271.8	.3	1	X
2476.5	.7	2	X
2681.2	.3	1	X
2885.9	0.0	0	
3090.6	.3	1	X
3295.3	0.0	0	
3500.0	.3	1	X
MISSING		8	
TOTAL		305	(204.71 = INTERVAL WIDTH)

Twenty is the minimum and 3500 the maximum value attained by variable 6 in the sample of respondents. The other values in the column titled Midpoint are the midpoints of the intervals for the histogram. The numbers in the column titled Hist% give the percentage of respondents falling in the corresponding interval. The column titled Count gives the actual number of respondents falling in the corresponding interval.

As can be seen from the histogram, the sample distribution of variable 6 is not symmetric. It is peaked at the left and has a long tail to the right. This shape suggests that the distribution of variable 6 may be approximated by the lognormal distribution.

Suppose that we wish to test whether variable 6 may be approximated by a lognormal distribution (we say approximated because there are only a finite number of lakeshore property owners in a county and a continuous distribution such as the lognormal can be at best an approximation to the actual discrete [finite] distribution of a variable). The Midas commands for this test are as follows:

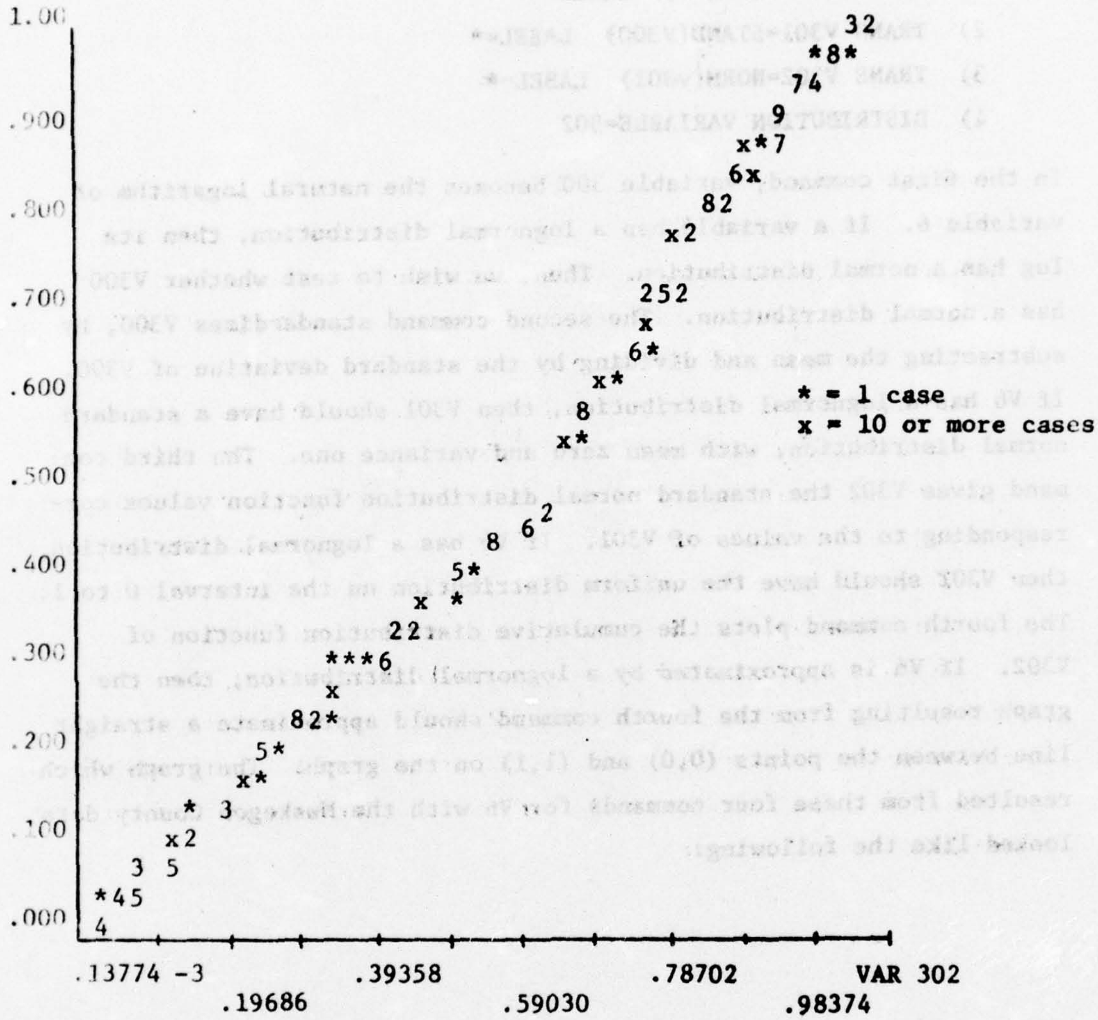
1) TRANS V300=LOG(V6) LABEL=*

2) TRANS V301=STAND(V300) LABEL=*

3) TRANS V302=NORM(v301) LABEL=*

4) DISTRIBUTION VARIABLE=302

In the first command, variable 300 becomes the natural logarithm of variable 6. If a variable has a lognormal distribution, then its log has a normal distribution. Thus, we wish to test whether V300 has a normal distribution. The second command standardizes V300, by subtracting the mean and dividing by the standard deviation of V300. If V6 has a lognormal distribution, then V301 should have a standard normal distribution, with mean zero and variance one. The third command gives V302 the standard normal distribution function values corresponding to the values of V301. If V6 has a lognormal distribution, then V302 should have the uniform distribution on the interval 0 to 1. The fourth command plots the cumulative distribution function of V302. If V6 is approximated by a lognormal distribution, then the graph resulting from the fourth command should approximate a straight line between the points (0,0) and (1,1) on the graph. The graph which resulted from these four commands for V6 with the Muskegon County data looked like the following:



DISTRIBUTIONAL ANALYSIS

CUMULATIVE SAMPLE DISTRIBUTION OF VAR 302 N = 297

A plot such as this one for variable 6 indicates that the variable being considered has a distribution which is well approximated by a lognormal distribution.

Each of the ten variables--assessed value, Propdepth, Propworth, total damage, total cost, Bluffheight, Beach depth, Blufflost, Bluffdist, and Beachlost--produced a reasonably straight line when these four commands were carried out for all six counties. (The one exception is that no assessed values were obtained for Huron County.)

TABLE 3.7 Summary Statistics for Lognormal Distributions and Proportions of Zero Values, Alcona County

Variable	$\hat{p}(0)$	Mean(log)	SD(log)
V6, Propdepth	34/552	5.7302	.75742
V18, Propworth	179/552	9.9268	.72008
Total damage	392/552	6.9772	1.4450
Total cost	379/552	7.1697	1.4883
V70, Bluffheight	316/552	1.9586	1.2188
V71, Beachdepth	330/552	3.1159	1.2184
V72, Buffloss	405/552	2.5361	1.2407
V73, Bluffdist	330/552	3.8789	.86709
V74, Beachloss	290/552	3.5658	.74262
V2*V4, Assessed value	7/552	9.7645	.69090

TABLE 3.8 Summary Statistics for Lognormal Distributions and Proportions of Zero Values, Chippewa County

Variable	$\hat{p}(0)$	Mean(log)	SD(log)
V6, Propdepth	16/605	5.7512	.83160
V18, Propworth	227/605	9.4468	.95947
Total damage	311/605	7.4099	1.3321
Total cost	276/605	7.6204	1.3204
V70, Bluffheight	212/605	1.9783	.78340
V71, Beachdepth	327/605	2.3921	.88893
V72, Bluffloss	288/605	2.0550	.91840
V73, Bluffdist	264/605	3.7797	.92954
V74, Beachloss	246/605	2.6963	.88805
V2*V4, Assessed value	3/605	9.0009	.75883

Note that since we cannot take the log of 0, zero values were ignored when performing this analysis. In Tables 3.7-3.12 below, $\hat{p}(0)$ gives the proportion of zero values in the sample for the indicated variables. Note that for variables such as total damage, zero includes both missing values and answers of zero damage. The questionnaire suggestions include a discussion of this problem. Also note that Tables 3.7-3.12 apply to distributions of variables for the samples of respondents and not for the counties themselves. In each case, no comparisons with the entire county could be made since values were not available for the entire county.

Also given in Tables 3.7-3.12 are the parameters of the lognormal distribution which approximates the distribution of the variable in the sample. These parameters are the mean and standard deviation of the log of each variable. For example, Mean(log) for variable 6 is the mean of log(V6) and SD(log) for variable 6 is the standard deviation of log(V6). However, $\hat{p}(0)$ for variable 6 indicates the proportion of responses of zero for variable 6 itself. Values are given to three significant figures.

Cost of protection is the sum of variables 46, 47, 51, 52, 57, and 58. No useful distribution could be found for cost of protection alone. However, when added to total damage the resulting variable called total cost has a distribution which is very well approximated by the lognormal distribution.

TABLE 3.9 Summary Statistics for Lognormal Distributions and Proportions of Zero Values, Huron County

Variable	$\hat{p}(0)$	Mean(log)	SD(log)
V6, Propdepth	13/280	5.5622	.78741
V18, Propworth	94/280	10.079	.81348
Total damage	154/280	7.2586	1.4176
Total cost	144/280	7.5930	1.4866
V70, Bluffheight	112/280	2.0876	.86493
V71, Beachdepth	148/280	2.7889	1.0327
V72, Bluffloss	149/280	2.6121	1.0258
V73, Bluffdist	123/280	3.9701	.84862
V74, Beachloss	124/280	3.2629	.82121
V2*V4, Assessed Value	280/280	-	-

TABLE 3.10 Summary Statistics for Lognormal Distributions and Proportions of Zero Values, Manistee County

Variable	$\hat{p}(0)$	Mean(log)	SD(log)
V6, Propdepth	8/204	5.8588	.99181
V18, Propworth	66/204	10.151	.85366
Total damage	85/204	8.0406	1.3569
Total cost	78/204	8.4542	1.2833
V70, Bluffheight	64/204	3.3395	.94452
V71, Beachdepth	91/204	2.7002	.85929
V72, Bluffloss	90/204	2.9605	.92753
V73, Bluffdist	85/204	3.7172	.94542
V74, Beachloss	80/204	3.7089	.81231
V2*V4, Assessed value	11/204	9.3225	.80142

TABLE 3.11 Summary Statistics for Lognormal Distributions and Proportions of Zero Values, Muskegon County

Variable	$\hat{p}(0)$	Mean(log)	SD(log)
V6, Propdepth	8/305	6.25	0.89
V18, Propworth	64/305	10.2	0.83
Total damage	104/305	8.06	1.36
Total cost	85/305	8.37	1.39
V70, Bluffheight	70/305	3.99	0.85
V71, Beachdepth	150/305	2.92	1.25
V72, Blufflost	105/305	3.27	0.73
V73, Bluffdist	91/305	3.90	1.29
V74, Beachloss	91/305	4.12	0.95
V2*V4, Assessed value	3/305	9.63	0.91

TABLE 3.12 Summary Statistics for Lognormal Distributions and Proportions of Zero Values, Schoolcraft County

Variable	$\hat{p}(0)$	Mean(log)	SD(log)
V6, Propdepth	9/134	6.1577	.92119
V18, Propworth	60/134	9.4095	1.0057
Total damage	116/134	7.1162	1.1660
Total cost	114/134	7.1483	1.1537
V70, Bluffheight	95/134	2.0771	.83715
V71, Beachdepth	107/134	2.7266	1.0771
V72, Bluffloss	110/134	2.3033	1.2307
V73, Bluffdist	107/134	4.1810	.91392
V74, Beachloss	100/134	3.4060	1.0743
V2*V4, Assessed value	36/134	8.7991	.96774

4.0 USE OF THE LOGNORMAL APPROXIMATION

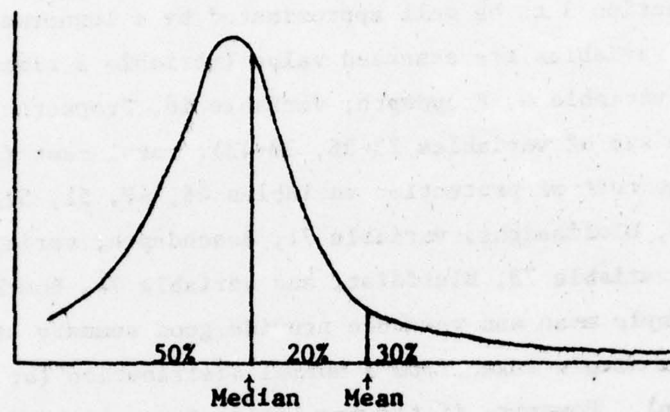
One might ask why we would want to describe any variables with lognormal distributional models. In this section we answer this question, show how to use these models, and consider distributional assumptions within individual reaches.

The variables discussed in this section are those which were found in Section 3 to be well approximated by a lognormal distribution. The variables are assessed value (variable 2 times variable 4, V2*V4); variable 6, Propdepth; variable 18, Propworth; total damage (the sum of variables 23-36, 38-43); total cost (total damage plus the cost of protection variables 46, 47, 51, 52, 57, 58); variable 70, Bluffheight; variable 71, Beachdepth; variable 72, Blufflost; variable 73, Bluffdist; and variable 74, Beachlost.

The sample mean and variance provide good summary statistics for a random sample taken from a normal distribution (or a near normal distribution). However, if the population from which we are sampling is skewed and/or heavy-tailed (with outliers) as are these variables (see Figure 4.1 below), then such statistics may be deceiving and we should be wary of conclusions based on them. Consider, for example, variable 74, Beachlost. Its mean is 70.17 feet. However, the median of this variable is 30.00 feet. The mean in fact lies between the 70.1 and 70.5th percentile of the questionnaire population. Although it is true that the average beachlost was 70.17 feet, 70% of the respondents lost less and 30% lost more. Let the graph in Figure 4.1 serve as an approximation to the histogram of variable 74, Beachlost. Then the median and the mean would appear in relative positions as shown. Note that the median is defined to be that value above which 50% of the population values lie and below which 50% of the population values lie. The percentages in Figure 4.1 indicate the approximate percentages of respondents in Muskegon County whose answers fell in the ranges 0-30.00 feet, 30.00-70.17 feet, and 70.17 feet or more. As can be seen from the figure, in the case of skewed variables such as these, the median serves as a better measure of"central tendency"

or "where the greatest proportion of the population lies" than does the mean. Reporting the sample median along with the sample mean and variance reveals this skewness in the population but not the shape of the distribution, for which a lognormal probability model is needed.

Figure 4.1 Comparison of the Relative Positions of the Median and the Mean for a Skewed (lognormal type) Distribution



Approximating the distributions of these ten variables with a lognormal model gives a more complete description of the variables than can be achieved by merely reporting the sample mean and variance of each variable. As discussed in the preceding paragraph, the sample mean and variance provide no clue as to the shape of the distribution of a variable. In fact, reporting only the mean and variance can be very misleading since many people think of the mean as being in the "middle" of the population, while Figure 4.1 indicates that this need not be the case. (The mean is actually in the "middle" of the population generally only for symmetric distributions such as the normal, when the mean and median coincide.) However, if a variable is well approximated by a lognormal distribution, as are these ten variables, then reporting the mean and variance (or the standard deviation, which is the square root of the variance) of the <u>log</u> of the variable completely describes the shape of the distribution. The reason is that the log of a lognormal variable has a normal distribu-

tion, and a normal distribution is completely specified by its mean and variance. Therefore, by reporting the mean and variance of the log of the variables rather than the mean and variance of the variable itself, we completely specifiy the distribution of the variable, but have still only used two numbers to do so.

The following examples illustrate some additional uses of the lognormal model not possible by reporting just the sample mean and variance.

Example 1

Let us suppose that our population of interest is the population of respondents for Muskegon County. The variable of interest at the moment is variable 6, Propdepth. We would like to take a 20% random sample of our population and construct an interval (L, R) based on this 20% random sample so as to predict that 90% of the property depths of the entire population of respondents will lie in this interval. Such an interval is called a tolerance interval and is constructed as described below.

Note that by taking a 20% random sample of the population of respondents in Muskegon County, we are simulating the proposed sampling scheme for future surveys described in the section on sampling. However, our 20% simulated sample is from the population of respondents, and as such, constructs tolerance intervals for the population of respondents. The proposed sampling scheme, on the other hand, is a 20% random sample of the entire county population. Tolerance intervals based on this proposed sampling scheme will give us predictive intervals for the entire county population.

For one 20% random sample of the respondent population, we obtained the following descriptive measure on log(V6):

Sample size = n = 61

Mean = $\hat{\mu}$ = 6.1552

Standard deviation = $\hat{\sigma}$ = .92182

We approximate the distirbution of log(V6) with a normal distribution

with mean $\hat{\mu}$ and variance $\hat{\sigma}^2$, denoted normal($\hat{\mu}$, $\hat{\sigma}^2$).

Classical normal theory tolerance intervals are constructed for log(V6) and are given as

$$(\hat{\mu} - \eta\hat{\sigma}\sqrt{1 + 1/n} \ , \ \hat{\mu} + \eta\hat{\sigma}\sqrt{1 + 1/n} \ ) \ *$$

where $\eta = \Phi^{-1}(.95)$ = 95th percentile of a normal (0, 1) = 1.65. For our example this interval is (4.626, 7.684). If we predict that 90% of the values of log(V6) are within this interval, then we also predict that 90% of the values of V6 will lie in $(e^{4.626}, e^{7.684})$ = (102.1, 2173) = (L, R). Such statements cannot be constructed from summary statistics like the mean and variance without an underlying probability model. A normal probability model applied directly to V6 would be totally inappropriate because of the skewed histogram for V6.

The numbers used in this illustrative example represent a Midas simulation of a 20% random sample of respondents to the mailed questionnaire for Muskegon County. The actual frequency of questionnaire respondents from Muskegon County with a property depth between 102.1 and 2173 is .92--a good tolerance interval!

Note that in computing the tolerance interval above, we used the percentile of a normal distribution, while Draper and Smith say to use the percentile of a student t-distribution with n-1 degrees of freedom, t(n - 1). When n is large (say, n > 30), the percentiles of the t(n - 1) distribution are very close to the percentiles of the normal (0, 1) distribution. Therefore, when n > 30, either the percentiles of the normal (0, 1) or the t(n - 1) may be used in constructing tolerance intervals.

Example 2

Suppose our population is the same as in Example 1 and we would

---

* Draper and Smith, Applied Regression Analysis, John Wiley & Sons, Inc. (1966), page 24.

like to estimate the fraction F of this population whose property depth is in a given interval, for example, the interval (120, 2000). If we consider the same 20% random sample as in Example 1, our log-normal model allows us to approximate the distribution of log(V6) as normal ($\hat{\mu}$, $\hat{\sigma}^2$). With this normal approximation to the distribution of log(V6), we estimate the fraction of the population within the interval (120, 2000) as follows: if X has a lognormal distribution with parameters $\hat{\mu}$ and $\hat{\sigma}^2$ (as we are using to approximate the distribution of V6), then logX has a normal distribution, normal ($\hat{\mu}$, $\hat{\sigma}^2$), and $\hat{F}$ = Prob $\{120 < X \leq 2000\}$ = Prob $\{\log(120) < \log X \leq \log(2000)\}$ = Prob $\{4.787 < \log X \leq 7.601\}$ estimates the fraction F of the population in the interval (120, 2000). Since logX has distribution normal ($\hat{\mu}$, $\hat{\sigma}^2$), this probability after standardization is

$$\hat{F} = \Phi\left(\frac{7.601 - \hat{\mu}}{\hat{\sigma}}\right) - \Phi\left(\frac{4.787 - \hat{\mu}}{\hat{\sigma}}\right)$$

where Φ is the standardized normal cumulative distribution function (c.d.f.). That is, if Z has a (standard) normal distribution with mean 0 and variance 1, normal (0, 1) and if z is any real number, then $\Phi(z) = P\{Z \leq z\}$ is the probability that Z has a value less than or equal to z.

For our example, $\hat{F}$ is then computed from normal tables to be $\hat{F}$ = .9418 - .0694 = .872. That is, our estimate of the fraction F of the population of Muskegon County respondents whose property depth falls in the interval (120, 2000) is F = .872. The actual fraction of respondents from Muskegon County with property depths falling in the interval (120, 2000) is F = .9, so we see that F as computed above is a good estimate of the true fraction F.

Tolerance intervals and estimates of fractions within intervals for a population based upon a random sample cannot be constructed from a sample mean and variance without an underlying probability model. When the mean and variance of the log of a variable which as an approximate lognormal distribution are reported, tolerance intervals such as in Example 1 and estimates of fractions within intervals such as in Example 2 are easily constructed.

What distributional assumptions can be made about these ten variables within individual reaches of Muskegon County? Midas sample distributions fitting lognormal distributions (as described on pages and) were plotted for each of the lognormal variables, stratifying on the reach number variable, V80. For each variable this has the effect of plotting five sample distributions--one based on each reach number level. The Midas commands to perform this for V6 are:

```
TRANS V200=LOG(V6)  LABEL=*
TRANS BYSTRATA V201=STAND(V200)  LABEL=*  STRATA=V80
TRANS V202=NORM(V201)  LABEL=*
DISTRIBUTION BYSTRATA VARIABLE=202  STRATA V80
```

Note the similarity between these commands and those on page . The difference is the BYSTRATA modifier in two commands which stipulates that these individual commands be independently performed for each stratum level specified by the level of V80. The results of this sequence of commands applied to each of the overall lognormal variables indicated good lognormal fits for most variables in strata whose sizes were not too small. Table 4.1 describes the lognormal fit in each stratum of Muskegon County for some of the lognormal variables. Stratum levels not mentioned fit well.

TABLE 4.1 Bystrata Lognormal Fits for Muskegon County

Variable Number	Variable Name	
V2*V4	assessed value	Bad fit, strata = 4
V6	Propdepth	O.K.
V18	Propworth	O.K.
	total damage	Marginal fit, strata = 4
	total cost	O.K.
V73	Bluffdist	O.K.
V74	Beachlost	Marginal fit, strata = 4

Stratum 4 was an especially small stratum and the only stratum not fitting well. Lognormal models within each reach for these variables seem to work reasonably well when the strata are large enough. Similar results were found to hold for the other counties considered. The suggested sampling scheme (to census those reaches which are small enough) will alleviate problems like those of stratum 4 since the actual distribution will be known when the entire stratum is surveyed.

A final note is in order concerning the use of the techniques discussed in this section. In future surveys, when random samples are taken, variables may be tested for approximate lognormal distributions, as described here and in section 3. For those variables well approximated by a lognormal distribution, tolerance intervals and estimates of fractions of the population falling in certain intervals may be obtained as described in this section. When a random sample has been taken, these tolerance intervals should provide very good "interval estimates" of the values of the variables within the entire population. Similarly, an estimated fraction should provide a very good estimate of the actual fraction of the entire population lying within an interval. Such intervals and estimates are only good, however, when a random sample has actually been taken so that the sample may be considered representative of the population as a whole. The section on proposed sampling schemes has a discussion of this problem of obtaining a true random sample.

5.0 OUTLIERS

Outliers are isolated values of a variable which are much larger or much smaller than the vast majority of other values, or values which seem spuriously extreme. An outlier in coded data may have one of the following sources:

1) It is a legitimate extreme value correctly given by the respondent and correctly coded by the coder;
2) It is an incorrect response by the respondent which was then coded by the coder;
3) It is not an extreme value, but a mistake was made in the coding process.

Extreme values can have large effects on the sample statistics for the data. For example, suppose that we have six responses for question B3 of the mailed questionnaire, a dollar amount for total damage suffered from erosion. Suppose that the responses given are 2000, 2500, 3000, 3100, 4000, 16000. The mean

$$\bar{x}_1 = 1/6 \sum_{i=1}^{6} x_i$$

of these six values is 5100 and the standard deviation

$$s_1 = \sqrt{1/5 \sum_{i=1}^{6} (x_i - \bar{x}_1)^2}$$

is approximately 5000. Note that all of the first five values fall below the mean $\bar{x}_1 = 5100$. On the other hand, the mean of the first five values

$$\bar{x}_2 = 1/5 \sum_{i=1}^{5} x_i$$

is 2920 and the standard deviation

$$s_2 = \sqrt{1/4 \sum_{i=1}^{5} (x_i - \bar{x}_2)^2}$$

of the first five values is approximately 700. Therefore, in describing and analyzing these responses, it is important to know whether 16000 is a legitimate dollar loss for one of the respondents or if it is a piece of misinformation.

Another example of possible misleading results is in the use of 0 values for the damage variables as currently coded. Some responses coded as 0 are missing values while others represent responses of 0 damages. Since it is not known which 0's correspond to missing values and which to 0 damages, either all 0's are included in any analysis, or all excluded. If all 0's are included and many represent missing values (when the nonrespondents really had damage), then the mean may underestimate the actual damage suffered. However, if all are excluded and many represent respondents who indeed suffered no damage, then the mean may very well overestimate actual damage suffered. Careful wording of questions and careful coding of responses will help alleviate this problem. Recommendations are provided in the section on questionnaire suggestions.

The following is the recommended way to deal with possible outliers. The first step is to look at the data and decide if the values look reasonable. For instance, the value of 11100 feet for frontage looked like an outlier to us, but was known to be a legitimate value by a person more familiar with the source of the data. If any values do seem unreasonably large or small, check for a coding error. It is possible, for instance, that the keypunch operator simply made an error in transcribing the data and such a mistake can easily be corrected. If no coding error can be found, check back with the respondent (make a follow-up call or visit to the person's house) to verify the validity of the response or replace it with a corrected response.

In summary, the process for checking possible outliers is as follows:

1) Ask if the response is reasonable.

   If yes, fine. If no:

2) Check for a coding error.

If there is one, correct it. If none:

3) Contact the respondent.

Verify or correct the response.

Table 5.1 below lists variables from the mailed questionnaire for Muskegon County with values we thought might be outliers. Values such as these should always be checked for reasonableness by people who are familiar with the data, following the three steps given above.

TABLE 5.1 Possible Outliers in Muskegon County

Variable	Value(s)
V5, Frontage	11100 feet
V6, Propdepth	Over 2000 feet
V19, Propdelt	Over $30000; negative values
V27, Flood E	$20000; over $10000
V31, FloodJ1	$20000; over $10000
V32, FloodJ2	$18000; over $10000
V33, Erode A	$20000; over $10000
V34, Erode B	$10000
V35, Erode C	Over $10000
V36, Erode D	$87000
V38, Erode E	$50000; over $10000
V39, Erode F	$50000; over $10000
V41, Erode H	$9600
V42, ErodeJ1	Over $10000
V43, ErodeJ2	Over $10000
V70, Bluffheight	Over 500 feet (over 300 feet?)
V71, Beachdepth	800 feet; over 500 feet
V72, Blufflost	99 feet; over 50 feet
V73, Bluffdist	Over 900 feet (over 500 feet?)
V74, Beachlost	750 feet; over 500 feet

6.0 MAILED QUESTIONNAIRE DATA VS. INTERVIEW DATA FOR MUSKEGON COUNTY

This section compares information from self-administered (mailed) questionnaires with data from personal interviews which were given to thirty-four of the respondents to the mailed questionnaire. The interview data should provide a check on the mailed questionnaire responses. An individual may give different answers to a comparable pair of questions on the two forms, even though the two questions ask the same thing. The data all comes from Muskegon County.

The first step in the analysis involves looking at the questions and seeing if they are really comparable. If the questions vary to an appreciable extent, the differences between the two sets of responses may be partly due to this variation and not just to the distinction between filling out a mailed questionnaire and being interviewed.

For property depth the appropriate mailed questionnaire item, question A5, is: "How many feet back from the present shoreline does your property extend (approximate depth)?" The corresponding interview item is question 24: "What is the total depth of this property?" Although the first question is clearer, the two questions should yield comparable responses.

For property worth the mailed questionnaire asks: "If you were to sell your property now, during high lake levels, how much do you think you could get?" The corresponding interview item is: "What is the market value of this property, given the high lake levels and existing rates of bluff erosion?" An individual would probably tend to give a lower value in response to the latter question since it includes the effects of erosion rates on property worth.

The questions on the variables measuring beach depth, beach loss, bluff height, bluff loss and distance from the bluff to the foundation of the residence are comparable, as the wording is essentially the same and the same diagram is used for reference in both questionnaires. The only discrepancy in these 5 sets of variables concerns bluff loss; the mailed questionnaire covers the time period from Labor

Day 1972 onwards while the interview is restricted to the two year period after this date.

The total damage figure from the personal interview is the sum of dollar amounts for questions 19 and 21. It involves damage to structure and contents of house, to structure and contents of other buildings, and to stairways, walls, and lawns; it also includes an "other" category and loss of rental income. Because of the ambiguity in the source of damage here, there are two possible sets of mailed questionnaire variables with which to compare this figure. One choice is all 20 variables in question B2; the other involves the 10 variables in B2 corresponding to erosion damage. The first choice involves damage from both flooding and erosion. The mailed questionnaire items constitute a more detailed list including damage to structure and contents of residence, to garages and outbuildings, docks, boathouses, stairways, ramps, grounds, landscaping, trees, clean up costs, septic system, loss of rental income, and other. The time periods are the same for the two sets of questions. Because of this and inclusion of the "other" category, the questions from the two forms are roughly comparable.

The questions involving cost of protective measures are set up quite differently. The mailed questionnaire data comes from B4c: "Has any protective action been taken (by you) with regard to your property since Labor Day, 1972? What did it cost?" (with breakdowns for materials and labor). The interview asks about "any protective action you have taken at any time to reduce the risk of damage to your property, due to high lake levels." The interview form asks for itemizations of physical relocation of buildings, temporary or emergency protective actions and permanent structural protection; the total cost of protection for the interview data is the sum of costs for these three. The two sets of questions again specify different time periods.

For both methods, total cost is the sum of total damage and cost

of protection.

Both forms ask whether any protective actions were taken. However, the questions vary substantially in meaning and in the information they elicit. The mailed questionnaire contains an item asking whether the respondent has taken any protective action, suffered any damage or is under risk of damage. The interview asks whether or not one or more of three modes of protective action was taken, a narrower definition. Also, from the coding of the interview data it is not possible to distinguish between people who took no action and persons who skipped the questions.

Comparisons of total damage from the mailed questionnaire with total damage from the interview were made using both sets of questionnaire damage variables (i.e., flooding <u>and</u> erosion damage variables). Because there were few responses to the flood damage questions on the mailed questionnaire, total damage using only erosion damage variables was not much different from total damage using erosion and flood damage variables. In this case, we arbitrarily decided to use all twenty damage variables (both erosion and flooding) from the mailed questionnaire to compute total damage for this analysis. This would not necessarily be appropriate for other counties where the ambiguity in interview questions with regard to the source of damage could lead to differing interpretations of the interview items by the person being interviewed, and where flooding may have caused significant damage to lakeshore property.

Thus, care should be taken when comparing mailed questionnaire and interview data for the variables property value, cost of protection and (to a lesser extent) bluff loss and total cost. Bear in mind that different questions may encourage higher or lower responses, depending upon the wording of the questions. A better evaluation of the two forms could be made if the wording were identical for comparable items on the two forms, as recommended in section 2 on questionnaire suggestions.

The following analysis of the two methods (mailed questionnaire vs. personal interview) is based on data from 34 individuals in Mus-

kegon County who returned the questionnaire and were also interviewed. The descriptive statistics of these 34 individuals were compared with the descriptive statistics of all 305 questionnaire respondents with respect to the following ten variables: property depth, property worth, beach loss, beach depth, bluff height, bluff loss, bluff distance, total damage, cost of protection, and total cost. The two groups were similar, indicating that the individuals who responded and were interviewed are representative of the wider class of all respondents. Therefore, the conclusions about the differences between questionnaire and interview responses based on this sample of 34 respondents should be applicable to the entire class of respondents.

Most of the following analysis will be concerned with the variables on the mailed questionnaire which are well approximated by a lognormal distribution. The cost of protection will also be considered. These are the most important variables and constitute most of the comparable variables which appear on both forms. The analysis includes the following: tests of significance of the differences between responses to the two forms, examination of trends in the differences and an evaluation of the two methods of obtaining information (i.e., mailed questionnaire vs. personal interview).

First, tests of significance were made for the lognormal variables. Recall that the natural logarithm of a variable with the lognormal distribution has a normal distribution. Thus, normal theory techniques can be used on the natural logarithms of these variables.

The fact that the same variables were covered by the two types of questionnaires and that the distributions of these variables are well approximated by the lognormal distribution indicates that a multivariate paired t-test might be appropriate on the differences of the logs of these variables. That is, we have nine approximately lognormal variables with values for both the mailed questionnaire and the personal interview (the ten listed above excluding cost of protection). We may consider for each variable the "pairs" consist-

ing of the response to the mailed questionnaire and the response to the personal interview for each of the thirty-four individuals who responded to both. The multivariate paired t-test is a technique for testing whether or not the means within each pair are equal for a group of variables (all with respect to the same group of individuals --in this case the thirty-four respondents). The problem with using this procedure is that it is based on the assumption that the logs of the variables together have a multivariate normal distribution, not just that the logs of the variables each have normal distributions separately. This assumption is hard to check. Also, there were insufficient observations to use this procedure on all nine variable "pairs" (i.e., there were not enough individuals who responded to <u>all</u> the relevant questions on both forms to make the test possible).

However, assuming an underlying joint multivariate normal distribution, the multivariate paired t-test was performed on a subset of the variables: property depth, property worth, beach loss, bluff loss and total cost. The hypothesis to be tested is that the means of the logs of these variables are the same for the mailed questionnaire and the personal interview. The test uses Hotelling's T^2 statistic, which is based upon the sample means, variances and covariances of the differences between the responses to the two types of questionnaires for the five variables being considered. (Covariance is a measure of a linear relationship between the values of two variables.) The value of Hotelling's T^2 statistic for this data is 4128, which is significant at any reasonable level. For instance, if we choose a level of 0.05, then the probability of obtaining a value for Hotelling's T^2 statistic as large or larger than 4128 when the paired means are really equal is less than or equal to 0.05. Therefore, if the underlying assumption of multivariate normality holds, we may reject the hypothesis that the mailed questionnaire responses and personal interview responses have equal means for the variables property depth, property worth, beach loss, bluff loss and total cost.

While the multivariate paired t-test tests whether the two

methods have the same means for several variables, the (univariate) paired t-test tests whether the two methods have the same mean for one given variable. For example, we may use the paired t-test to test the hypothesis that the average total damage figure from the mailed questionnaire is the same as the average total damage from the interview, using the log of the variables so that the normality assumptions of the procedure are satisfied. As the multivariate procedure could not be used for all nine of the lognormal type variables, paired t-tests were performed on each of them to test whether the logs of the answers to comparable questions have the same mean for the mailed questionnaire and personal interview. The statistic for the paired t-test is computed for a given variable in the following manner: let N represent the number of individuals who gave a non-zero answer for the variable for the mailed questionnaire and the interview and consider only these individuals. Let x_i represent the response of the ith individual on the mailed questionnaire item and $\hat{x}_i$ represent the response of the same person on the comparable interview item. Let $d_i = \log x_i - \log \hat{x}_i$ and $\bar{d}$ be the mean of these d's. Let

$$T = \bar{d}\sqrt{N}/s \qquad \text{where } s^2 = \frac{\sum_{i=1}^{N}(d_i - \bar{d})^2}{N-1}$$

Then T has the t-distribution with N - 1 degrees of freedom if the hypothesis is true. In Table 6.1, Signif, the attained significance level, is the probability of getting as extreme or more extreme a result by chance if the hypothesis is true. Thus, a small value of Signif indicates that the hypothesis is likely to be false. The second column of Table 6.1 gives the attained significance levels for the paired t-test on the logs of the nine approxiamtely lognormal variables. This information was obtained by giving a Midas command of the form

PAIR VAR=1,2

where 1 and 2 stand for the indices of the pair of variables to be compared using the paired t-test.

TABLE 6.1 Comparison of Questionnaire and Interview Data for Respondents of Both

Variable	Attained Signif Level of t-test	Attained Signif Level of Median Test
Property depth	.01	-
Property worth	.79	-
Bluff height	.15	-
Beach depth	.95	-
Bluff loss	.74	.33
Bluff distance	.88	-
Beach loss	.69	.43
total damage	.19	.40
total cost	.78	.69
cost of protection	-	.45

Except for property depth, the attained significance levels for the paired t-tests are fairly large. Thus, except for property depth, the univariate paired t-tests do not indicate a significant difference between the means of these nine variables for the mailed questionnaire and the personal interview. (A dash in a column of Table 6.1 for a variable indicates that the corresponding test was not performed for that variable.)

As noted in the questionnaire section, a value of zero may correspond to either a missing value or a data value of zero; the log procedures above rule out all values of 0, as 0 does not have a finite log. Throwing out all values of zero for the variables corresponding to losses and dollar expenditures could lead to incorrect conclusions. Therefore a less powerful technique, the median test, was used on some of these variables themselves, rather than their logs. The median test includes values of zero in the analysis. The median test only looks at the sign of the difference between the questionnaire response and the interview response; it makes no assumption about the distribution of the variables.

The values in the third column of Table 6.1 came from using the median test. The appropriate Midas command to perform the median test is

```
RPAIR VAR=1,2
```

where 1 and 2 stand for the indices of the pair of variables to be

compared using the median test. As indicated by the significance levels attained by the median test for those pairs of variables tested (listed in the third column of Table 6.1), the median test suggests that there are no appreciable differences between the means of the variables tested for the mailed questionnaire and the personal interview.

Recall that the multivariate paired t-test indicated that there were significant differences between the means of the variables tested for the questionnaire and the interview, while the univariate paired t-test and the median test indicated that there were not (for those variables tested). Because of the assumption required by the multivariate paired t-test of joint multivariate normality of the variables tested, an assumption which may very well not be satisfied here, we must view the results of the multivariate paired t-test with suspicion.

Although the univariate paired t-test and the median test indicate no significant differences between the means of the variables for the mailed questionnaire and the personal interview, the chance of a type II error may be substantial in view of the sample size (and lack of positive correlation between the responses). That is, the probability that the hypothesis of no difference will be accepted when a difference, in fact, exists may be high. Examination of trends in the data leads us to the tentative decision that significant differences may exist, and would be found with additional data.

Recall from the discussion in sections 3 and 4 that for variables with an approximate lognormal distribution, the mean is a poor measure of central tendency, being located rather far out on the right tail of the distribution (see Figure 4.1). A better measure of central tendency for approximate lognormal variables may be obtained as follows. Suppose we are considering total damage, which has an approximate lognormal distirbution. Find the mean of the log of total damage and then raise "e" to the power of this mean. Midas commands

may be used to perform these computations. Suppose total damage for the mailed questionnaire is V100 and for the personal interview is V101. Then the Midas commands are:

```
TRANS FUNCTION=LOG  VAR=100,101  RESULT=500,501
DESCRIBE  VAR=500,501
```

The TRANS command computes V500 equal to the log of V100 and V501 equal to the log of V101. The DESCRIBE command prints out the mean of V500 and of V501. Then we compute exp(mean of V500) and exp(mean of V501) where exp(·) means "e" raised to the power inside the parentheses. For example, the mean of the log of total damage (questionnaire) is 8.758 and the mean of the log of total damage (interview) is 7.7078. Raising "e" to the mean of the logs we obtain

$$e^{8.758} \simeq 6,360$$

$$e^{7.7078} \simeq 2,230$$

These measures of central tendency for mailed questionnaire and personal interview for the nine approximately lognormal variables appear in Table 6.2.

TABLE 6.2 Comparison of Measures of Central Tendency, $\hat{m}$

Variable	Questionnaire	Interview
Property depth	413	479
Property worth	23,200	20,900
Bluff height	59.7	40.6
Beach depth	11.9	17.1
Bluff loss	25.0	19.5
Bluff distance	56.0	48.3
Beach loss	51.4	44.2
total damage	6,360	2,230
total cost	5,820	5,120

Let us call the measure listed in Table 6.2 for a variable $\hat{m}$ and call the mean for the variable $\hat{\mu}$. Call the mean of the log of the variable $\hat{\mu}$(log). Because the log of the variable is approximately normally distributed, $\hat{\mu}$(log) will appear near the center of the distribution of the log of the variable, near where the graph reaches

its highest point. Therefore, $\hat{m} = \exp[\hat{\mu}(\log)]$ will be near where the graph of the variable itself reaches its highest point--closer to the median than to the mean (see Figure 4.1). Since it is around this peak in the distribution that the greatest percentage of the variable's values lie, we see that m as listed in Table 6.2 really is a better measure of "central tendency" or "near where the greatest percentage of values lie" than is the mean.

Comparing the values in columns 2 and 3 of Table 6.2, we see that these thirty-four respondents in general gave higher answers on the mailed questionnaire than during the interview for the variables property worth, bluff height, bluff loss, bluff distance, beach loss, total damage and total cost. Thus, respondents gave higher figures on the mailed questionnaire for all four damage and loss variables (bluff and beach loss, total damage and cost) than they did for the personal interview. Those who wish to use this data to make decisions should keep this finding in mind. If people's responses in a personal interview setting are considered closer to their true losses than their responses to a mailed questionnaire, then the mailed questionnaire responses for damages and losses are almost consistently "inflated." Of course, without true values against which to compare responses, we have no way of knowing which set of responses (mailed questionnaire or interview) are more reliable. Also, one cannot rule out the very definite possibility that differences in responses are due to differences in wording of the questions in the two settings. As noted in section 2, this problem of interpretation would be eliminated if comparable questions were worded identically on both forms.

The substantial difference between the mailed questionnaire and interview values of m for total damage led to further investigation of the relationship between these two measures. A scatter plot of total damage (questionnaire) against total damage (interview) indicated a definite linear relationship. Various regression models were tried on the log of total damage (questionnaire) and the log of total damage (interview).

A regression model states that the value of some (dependent) variable is a specified function of one or more other variables (the inde-

pendent variables), plus a random error term. The error term is assumed to have a normal distribution with mean zero and constant variance σ^2; errors on different trials are assumed to be independent. Logs were used here because of the regression assumption that the errors have a normal distribution.

A linear regression of the log of total damage (interview) on the log of total damage (questionnaire) was performed. Let Y stand for the log of total damage (interview) and X for the log of total damage (questionnaire). Then the model is:

$$Y = \beta X + \text{error}.$$

We wish to find b, the least squares estimate of β. Using the Midas REGRESSION command, b was found to be approximately .92. The Midas command has the following form:

REGRESSION VAR=1,2 OPTION=MEANZERO

where V1 is the dependent variable, log of total damage (interview), and V2 is the independent variable, log of total damage (questionnaire). Setting OPTION equal to MEANZERO forces the y-intercept in the regression to be zero. Although the hypothesis that $\beta = 0$ can be rejected, this regression model does not explain very much of the variation in the log of damage (interview), since R^2 is only .00147 (where R^2 is the fraction of variation in Y which is explained by a linear relationship between Y and X). Other regression models were fit to the data. In all cases R^2 was small and/or the regression was not significant. Thus, while log of damage (interview) tends to increase with log of damage (questionnaire), there are other factors which cause most of the variability in log of damage (interview). This lack of a clear-cut relationship between the logs of damage (interview) and damage (questionnaire), and therefore between damage (interview) and damage (questionnaire), indicates that the respondents' answers for total damage on the mailed questionnaire and in the interview setting are not very consistent with each other.

Finally, we make an evaluation of questionnaire data and interview data with respect to the variables discussed above. The presence of the interviewer has a conservative effect on responses to questions

about total damage and total cost. The same is true to a lesser extent for beach loss and bluff loss. As the interviewer could make observations while on the property, the interview data might be considered more reliable than questionnaire data with respect to these variables.

For property worth, the same remarks hold as in the previous paragraph. However, we have the additional information of assessed value for these properties with which to compare the respondents' estimates of property worth. As these variables are approximately lognormal, the comparison will be made by taking logs, then computing the mean of the logs and then raising "e" to the power of the mean of the logs; that is, by computing $\hat{m}$'s as we did for Table 6.2. This information is contained in Table 6.3.

TABLE 6.3 Comparison of Measures of Central Tendency, $\hat{m}$

Variable	Mean of Logs
Property worth (questionnaire)	23,200
Assessed value	14,200
Property worth (interview)	20,900

The means for both questions involving a personal opinion about property value are above the mean for the assessed value of the property; however, the interview data comes closer to the assessed value than does the questionnaire item. Which of these values is to be considered most reliable must be decided by those who will use the data.

Finally, for property depth, questionnaire responses tend to have lower values than do interview responses.

7.0 RESPONDENTS VS. NON-RESPONDENTS IN MUSKEGON COUNTY

As already noted in section 6, thirty-four respondents to the mailed questionnaire in Muskegon County were also given a personal interview. In addition, a random sample of fourteen nonrespondents to the mailed questionnaire (including at least two nonrespondents from each of the five reaches in Muskegon County) was selected, and these fourteen nonrespondents were also given a personal interview. In this section we compare the answers given by these respondents and nonrespondents in the personal interview setting. This comparison is important since if nonrespondents vary substantially in their answers (especially for the damage and loss variables) from respondents, then figures based only on respondent replies to the mailed questionnaire will not be representative of the entire population of lakeshore property owners.

Hypothesis testing and descriptive measures are used to examine and compare the ten variables property worth, property depth, beach loss, beach depth, bluff loss, bluff height, bluff distance, total damage, cost of protection and total cost for respondents and nonrespondents.

From section 3 on distributions we know that the variables listed above, with the exception of cost of protection, are approximately lognormal for questionnaire respondents. There are not enough cases to get a distribution for these variables for nonrespondents. For the moment, suppose that these ten variables have the same type of distribution for both respondents and nonrespondents.

If the logs of the nine variables which are approximately lognormal (cost of protection is excluded) had a joint multivariate normal distribution, then a multivariate analysis of variance would be appropriate to test whether or not respondents and nonrespondents have the same set of means for these nine variables. There were not enough cases to perform a multivariate analysis of variance on the logs of all nine variables. Instead, the procedure was used on the logs of a subset of the variables: property worth, property depth, and total cost. The hypothesis to be tested is whether the (logs of)

answers given by respondents to the personal interview have the same means for these three variables as do the (logs of) answers given by nonrespondents. The attained significance level when the multivariate analysis of variance was performed on the logs of these three variables was .3849. That is, according to this procedure, if the means for these variables were really the same for respondents and nonrespondents, then the probability of seeing differences between the means as great as or greater than those for this random sample of respondents and nonrespondents is .3849. Therefore, the multivariate analysis of variance indicates no significant differences between the means of (the logs of) property worth, property depth and total cost for respondents and nonrespondents.

As with the multivariate paired t-test discussed in section 6, however, the assumption of joint multivariate normality for the logs of these approximately lognormal variables may very well not hold. Therefore, the results of this multivariate analysis of variance should be viewed with caution.

The (univariate) two-sample t-test, which should be distinguished from the paired t-test used in section 6, may be used to test whether respondents and nonrespondents have the same mean for one given variable. (With the paired t-test we tested equality of means for two different responses given by the same individual--such as interview total damage vs. questionnaire total damage for each respondent. With the two-sample t-test we test for equality of means of the same variable for two different individuals--such as interview total damage for respondent vs. interview total damage for nonrespondent.) For example, we may use the two-sample t-test to test the hypothesis that the average total damage for respondents is the same as the average total damage for nonrespondents, using the log of the variable so that the normality assumptions of the procedure are satisfied. Two-sample t-tests were performed on the logs of each of the nine lognormal type variables to test whether the logs of these variables have the same mean for respondents and nonrespondents.

The statistic used for the two-sample t-test is computed in the

following manner. Let X_{ij} represent the log of the variable X for the jth individual in the ith group (i is either 1 or 2). Let $\overline{X}_1$ represent the mean of the logs for the individuals in group 1 (in this case respondents) and $\overline{X}_2$ represent its counterpart for the second group (in this case nonrespondents). Let N_1 stand for the number of individuals who belong to the first group (respondents) and let N_2 be the number of individuals in the second group (nonrespondents), where we count only cases with positive values, so the log is defined. Let

$$s_i^2 = \sum_{j=1}^{N_i} (X_{ij} - \overline{X}_i)^2 / (N_i - 1)$$

be the sample variance for the ith group i=1,2 and let

$$s_p^2 = \frac{(N_1 - 1)\, s_1^2 + (N_2 - 1)\, s_2^2}{N_1 + N_2 - 2}$$

be the pooled variance. If the two groups have the same means and the same variance for the variable X, then it can be shown using distribution theory that

$$T = \frac{\overline{X}_1 - \overline{X}_2}{s_p \sqrt{1/N_1 + 1/N_2}}$$

has a t distribution with $N_1 + N_2 - 2$ degrees of freedom.

The Midas command for a two-sample t-test has the following form:

```
STUDENT VAR=1 STRATA=V402
```

where variable 1 is the variable of interest (for example, total damage). The strata keyword defines the two groups which are being compared. For example, V402 has the value 1 for respondents and the value 2 for non-respondents.

Column 2 of Table 7-1 lists the attained significance levels achieved when the two-sample t-test was performed on the logs of the indicated variables. A dash in a column for a variable means the test was not performed for that variable.

The two-sample t-test is based on the assumption that the variance of the variable for the first group (respondents) is equal to the variance of the variable for the second group (nonrespondents). Column 3 of Table 7-1 lists the attained significance levels for each variable when a test of equal variances of the log of the variable for the two groups was performed. A small level of significance in column 3, say less than .1, indicates that the log of the variable probably does not have the same variance for respondents as for nonrespondents and that the two-sample t-test was therefore inappropriate. With this criterion, we should ignore the results of the two-sample t-test for property depth and beach loss, which correspond to .0931 and .0886, respectively, in column 3 of Table 7.1. The statistics for bluff height were not listed bacause of unequal variances.

TABLE 7.1 Tests for Significant Differences Between Respondents and Nonrespondents

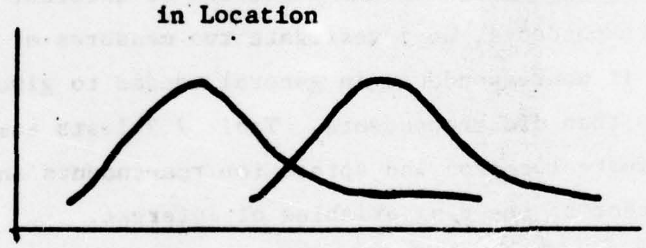
Variable	Signif t-test (on logs)	Signif test for equal var	Signif W.R.S. 0's inc	Signif W.R.S. 0's exc	Signif Median 0's inc	Signif Median 0's exc
Property worth	.1430	.3807	-	.1356	-	.2690
Property depth	.4360	.0931	-	.4809	-	.6060
Beach loss	.2266	.0886	.2890	.3359	.5618	.472
Beach depth	.2788	.2348	-	.44	-	.8242
Bluff loss	.0817	.4344	.1217	.1302	.3412	.3034
Bluff height	-	-	-	.6996	-	.7516
Bluff distance	.7299	.1649	-	.82	-	.484
Total damage	.9864	.2482	.3584	.8105	.3412	1.
Cost of protection	-	-	.5735	.28	1.	.4448
Total cost	.4120	.1294	.5326	.6060	.7516	.4578

The variables which passed the test for equal variances (using the criterion given above) are property worth, beach depth, bluff loss, bluff distance, total damage and total cost. The results of the two-sample t-test for each of these variables indicate that the mean of the log of each variable is the same for respondents and nonrespondents.

Two less powerful procedures were also used, making it possible to include values of 0 in the analysis. These procedures are the Wilcoxan rank sum test and the median test. Both test one variable at a time, as does the two-sample t-test. Both procedures were used on the variables themselves rather than on their logs.

The Wilcoxan rank sum procedure assumes that the distribution of a variable for respondents is the same as the distribution of the variable for nonrespondents, except that it may be "located" at a different point along the x-axis. An example is shown in Figure 7.1.

Figure 7.1 Two Distributions Alike Except for a Difference in Location



Note that without complete data for all respondents and nonrespondents in Muskegon County, we cannot test this assumption that the distribution of a variable is the same for respondents and nonrespondents except possibly for location.

The Wilcoxan rank sum procedure tests the hypothesis that there is no difference in location between the distributions--i.e., that the variable has the same distribution for both respondents and nonrespondents. Column 4 of Table 7.1 lists the attained significance levels for the Wilcoxan rank sum test when values of 0 are included. Column 5 gives the significance levels when values of 0 are excluded.

The median test tests the hypothesis thatthe samples of respon-

dents and nonrespondents come from populations having the same median for the variable being tested. Columns 6 and 7 of Table 7.1 give the attained significance levels for the median test, with values of 0 included and excluded, respectively.

The Midas command which provides results for both the Wilcoxan rank sum test and the median test has the form:

TWOSAMPLE VAR=1 STRATA=V402

where, as with the STUDENT command for the two-sample t-test, 1 stands for the index of the variable to be tested and the strata keyword defines the two groups to be compared (in this case, respondents and nonrespondents).

Both the Wilcoxan rank sum test and the median test indicate no significant difference between the means of the variables tested for respondents and nonrespondents.

Even though no statistically significant differences were found between the means of the variables of interest for respondents and nonrespondents, we investigate two measures of central tendency to see if nonrespondents in general tended to give higher or lower answers than did respondents. Table 7.2 lists some statistics which indicate location and spread for respondents and nonrespondents with respect to the ten variables of interest.

Two methods of comparison were used to look at possible differences in location between respondents and nonrespondents. The first method looks at $\hat{m}$, which is "e" raised to the mean of the log of a variable. As discussed in section 6, $\hat{m}$ comes closer to describing a "typical" value for a variable which is approximately lognormal than does the mean. The second method looks at the arithmetic mean $\hat{\mu}$ of the variable. The mean $\hat{\mu}$ for respondents and for nonrespondents was computed for cost of protection, which is not approxiamtely lognormal, and for beach loss, bluff loss, total damage and total cost. The mean is used for the last four variables because taking logs excludes 0 values and these four variables could take on 0 as a nonmissing value. The mean $\hat{\mu}$ has the disadvantage of being more sensitive to outliers

than is $\hat{m}$, and the merit of including 0 values depends on how many of the coded 0's correspond to answers of 0 (rather than to missing values).

Table 7.2 also gives the sample standard deviation s of each variable for respondents and for nonrespondents.

TABLE 7.2 Measures of Central Tendency and Standard Deviations for Respondents and Nonrespondents

Variable	$\hat{m}$	$\hat{\mu}$	s
Property depth			
Respondent	479	-	568
Nonrespondent	368	-	698
Property worth			
Respondent	20,900	-	25,000
Nonrespondent	10,700	-	18,400
Beach loss			
Respondent	44.2	48.0	46.9
Nonrespondent	61.9	62.0	47.1
Beach depth			
Respondent	17.1	-	26.2
Nonrespondent	10.4	-	16.0
Bluff height			
Respondent	40.6	-	39.6
Nonrespondent	47.0	-	28.3
Bluff loss			
Respondent	19.5	24.6	23.0
Nonrespondent	33.4	43.8	42.4
Bluff distance			
Respondent	48.3	-	169
Nonrespondent	61.1		105
Total damage			
Respondent	2,240	4,470	9,910
Nonrespondent	2,200	2,920	3,240
Cost of protection			
Respondent	-	3,500*	11,000
Nonrespondent	-	2,450	3,760
Total cost			
Respondent	5,120	7,970	20,910
Nonrespondent	3,120	5,370	7,000

* If the two large outliers are deleted, the cost of protection for respondents has a mean of 980.

As can be seen from Table 7.2, respondents tended to give larger answers for property worth, property depth, beach depth and total cost. Nonrespondents tended to give slightly higher answers for bluff height and bluff distance. Nonrespondents also tended to report greater beach and bluff losses. If all data values are used, respondents have a higher average expenditure for protection. However, if the one outlier of $59,300 for a respondent is discarded, the pattern is reversed giving nonrespondents a greater average cost of protection.

Thus, the figures in Table 7.2 indicate that nonrespondents in general have smaller, less valuable properties and have suffered greater beach and bluff losses from the high waters.

From column 4 of Table 7.2 we see that nonrespondents' answers fluctuate more for property depth and bluff loss. Respondents' answers fluctuate more for property worth, beach depth, bluff height, bluff distance, total damage, cost of protection, and total cost.

There are numerous variables which have a multinomial distribution, meaning that an answer falls into one and only one of a finite number of categories. However, with so few nonrespondents and so many missing values, it is not reasonable to compare respondents and and nonrespondents with respect to these multinomial type variables.

8.0 ESTIMATES OF POPULATION TOTALS FOR SIX COUNTIES

This section describes our search for appropriate estimates of total damages actually suffered in Alcona, Chippewa, Huron, Manistee, Muskegon, and Schoolcraft Counties. The four variables to be estimated will be called Damage, Cost, Blufflost, and Beachlost.

Damage is the sum of all losses due to flooding and erosion as listed by the respondents in question B2 of the questionnaire. Cost is the sum of Damage plus the cost of labor and materials for protective actions, as listed in B4 of the questionnaire. Blufflost and Beachlost are the answers to D3 and D4.

Our goal is to find and apply the best methods for estimating actual strata-wide totals of these four variables to the various counties, where the strata correspond to reach number classifications.

The data consists of responses to the mailed questionnaire for each county. Estimates for Muskegon County, however, may make use of the additional data found in the personal interviews.

The preliminary version of this report focused on ratio estimation as a means for extrapolating from the data to estimates of population totals for the variables above. Such extrapolation, however, requires an instrument variable, that is, auxiliary information available for the entire population which is highly correlated with the variable of interest. Unfortunately, no acceptable instrument variables were found for the variables above so that other extrapolation procedures had to be considered. Three alternative methods for extrapolation are presented below.

There are two ways to consider the mailed questionnaire data for constructing estimates of total damages. We first present these two methods for all the counties and then propose a third method applicable only to Muskegon County since it relies on the personal interview data.

In the first method of estimation, we assume that nonrespondents to the mailed questionnaire suffered negligible damage of the types we are interested in. Therefore, we use the total damages listed by

the respondents as estimates of the total damages suffered by all lakeshore property owners for each of the counties. These totals are listed in Table 8.1 and represent estimates of total county damages if nonrespondents are assumed to have suffered no damage. Since only 15% of Huron County was surveyed, the respondent totals in Table 8.1 for this county have been divided by .15 to give respondent estimates for the entire county.

TABLE 8.1 Total for Respondents to Mailed Questionnaires*

County	Reach	# Respondents	Damage	Cost	Blufflost	Beachlost
Alcona	1	431	333,000	436,000	3000	10,100
	2	60	30,000	39,000	360	1,200
	3	28	19,000	64,000	180	400
Total		(519)	(382,000)	(539,000)	(3500)	(11,600)
Chippewa	1	2	5,000	5,000	0	0
	2	9	19,000	21,000	40	50
	3	227	547,000	908,000	1880	4,850
	4	44	42,000	50,000	300	500
	5	0	0	0	0	0
	6	312	431,000	536,000	1310	2,800
	7	11	59,000	61,000	70	90
Total		(605)	(1,103,000)	(1,580,000)	(3600)	(8,300)
Huron	1	78	931,000	1,896,000	4300	13,600
	2	26	320,000	557,000	3900	5,300
	3	77	537,000	771,000	5700	11,300
	4	83	861,000	1,009,000	3900	4,800
	5	16	45,000	70,000	1300	5,600
Total		(280)	(2,700,000)	(4,300,000)	(19,000)	(41,000)
Manistee	1	54	137,000	204,000	790	1940
	2	7	35,000	36,000	240	390
	3	11	14,000	14,000	100	160
	4	19	33,000	70,000	430	940
	5	51	214,000	306,000	1020	1460
	6	8	12,000	16,000	20	80
	7	40	259,000	324,000	270	1120
	8	13	85,000	169,000	310	770
Total		(203)	(788,000)	(1,140,000)	(3200)	(6800)

TABLE 8.1 (cont'd)

County	Reach	# Respondents	Damage	Cost	Blufflost	Beachlost
Muskegon	1	43	337,000	470,000	950	2300
	2	112	537,000	982,000	2900	7700
	3	58	268,000	331,000	1000	4700
	4	25	51,000	68,000	500	1500
	5	49	129,000	215,000	940	4400
Total		(287)	(1,322,000)	(2,066,000)	(6290)	(20,600)
School-craft	1	49	30,700	34,000	310	670
	2	46	8,500	11,400	120	470
	3	9	1,000	1,000	10	40
	4	23	400	7,500	60	670
	5	2	0	0	0	0
Total		(129)	(41,000)	(47,000)	(500)	(1800)

* Figures are rounded off; Huron County entries have been divided by .15.

In the second method of estimation, we assume that respondents and nonrespondents are really no different in terms of damages suffered. In this case we consider respondents as a representative "sample" of the total population of lakeshore property owners and use responses to the mailed questionnaire to obtain estimates of county totals. We are not advocating this approach. It is in general a gross mistake to consider respondents and nonrespondents as being alike. Usually nonrespondents have reasons for not returning the questionnaires which make them very different from respondents. In no way should respondents be considered a random sample of all the lakeshore property owners. This approach is demonstrated below by estimating total damages for Muskegon County followed by estimates for the other five counties.

For Muskegon County the questionnaire respondents represent a 60% overall "sample" of the total county population. We would like to find good strata-wide estimates of totals for the four variables of interest. To do this we "simulate" the estimation procedure as follows. Suppose we take all questionnaire respondents as our popu-

lation and "simulate" 60% response to the questionnaire with a 60% stratified random sample (explained below) of the respondents. Then, we may compare response total estimates based on the 60% sample with actual response totals to establish the "best" type of estimator to use. This "best" estimator type can then be applied to all the respondent data to predict totals for the entire county population. This "best" estimator type, however, relies on the assumption that respondents and nonrespondents are really no different in terms of damage suffered by each.

To carry out this "simulation" for Muskegon County, two random samples of the population of respondents were drawn, such that each random sample consisted of approximately 60% of the members of the strata with reach number 1, and so on for each of the five strata. This process was repeated to obtain a second 60% random sample, different from the first.

Before explaining the types of estimators considered for estimating totals, we need some notation. For a given variable, let $\hat{x}_{ij}$ stand for the mean of that variable in the jth strata for the ith random sample. Suppose we are considering the variable Damage. Then for the first random sample, for instance, $\hat{x}_{11}$ stands for the mean of the Damage values in the first strata (reach number 1) of the first random sample; $\hat{x}_{12}$ stands for the mean of the Damage values in the second strata (reach number 2) of the first random sample; $\hat{x}_{15}$ stands for the mean of the Damage values in the fifth strata (reach number 5) of the first random sample. Similarly for the other random sample.

Again, for a given variable, let $\hat{x}_i$ stand for the grand mean for that variable in the ith random sample. If we are considering the variable Damage, then $\hat{x}_1$ stands for the grand mean of the Damage values for all respondents in the first random sample and $\hat{x}_2$ stands for the grand mean of the Damage values for all respondents in the second random sample. Thus, the strata are ignored and the entire random sample used to compute the grand means $\hat{x}_i$.

Finally, for a given variable, let $\bar{x}_j$ stand for the mean for that variable within the jth strata of the original population of all respondents. The $\bar{x}_j$'s are the actual strata means of the respondent population.

The respondent population in Muskegon County for the four variables we are considering consists of N = 287 respondents in K = 5 strata. There are $N_1 = 43$ respondents in strata 1, $N_2 = 112$ in strata 2, $N_3 = 58$ in strata 3, $N_4 = 25$ in strata 4 and $N_5 = 49$ in strata 5.

We are trying to estimate total losses within the strata of the respondent population. Therefore, we are going to try to estimate $N_1\bar{x}_1$, $N_2\bar{x}_2$, $N_3\bar{x}_3$, $N_4\bar{x}_4$ and $N_5\bar{x}_5$ for each of the four variables.

Let y_{ij} be the estimate of $\bar{x}_j$ for the ith sample. Then $N_i y_{ij}$ is used to estimate $N_j\bar{x}_j$. For example, y_{24} is the estimate of the mean $\bar{x}_4$ of the fourth strata of the respondent population, obtained from the second random sample. $N_4 y_{24}$ is the second random sample estimate of the fourth strata population total $N_4\bar{x}_4$.

There were eight types of estimators tried for each of the four variables. These types are tabulated in Table 2 below. Types 7 and 8 happen to be the same for Muskegon County, but will be different in some later counties.

TABLE 8.2

Type	Estimate of $\bar{x}_j$
1	$y_{ij} = x_{ij}$
2	$y_{ij} = x_i$
3	$y_{ij} = (1/2)\ x_{ij} + (1/2)\ x_i$
4	$y_{ij} = (1/4)\ x_{ij} + (3/4)\ x_i$
5	$y_{ij} = (3/4)\ x_{ij} + (1/4)\ x_i$
6	$y_{ij} = (N_j/N)\ x_{ij} + (1 - N_j/N)x_i$
7	$y_{ij} = (1/5)\ x_{ij} + (4/5)\ x_i$
8	$y_{ij} = (1/K)\ x_{ij} + (1 - 1/K)\ x_i$

Type 1 estimates respondent population strata means with sample strata means. Type 2 ignores strata differences and uses sample grand means to estimate respondent population strata means.

Types 3-8 each use some sort of average of sample strata and grand means to try and achieve better estimates of population strata means than can be obtained with Types 1 and 2 alone. Since some sort of average of $\hat{x}_{ij}$ and $\hat{x}_i$ uses more information than using either alone, it was expected that Types 3-8 would provide better estimates than Types 1 and 2. Sometimes this proved to be the case, as will be seen later.

Type 3 assigns equal weights of 1/2 to each of $\hat{x}_{ij}$ and $\hat{x}_i$. Types 4, 5 and 7 assign weights of 1/4, 3/4; 3/4, 1/4; and 1/5, 4/5, respectively, to $\hat{x}_{ij}$ and $\hat{x}_i$. Type 6 weights $\hat{x}_{ij}$ by the relative size of strata j in the respondent population, and Type 8 weights $\hat{x}_{ij}$ according to the number of strata in the county. For instance, for Type 6 $y_{13} = (58/287)\ \hat{x}_{13} + (1 - 58/287)\ \hat{x}_i$.

A discussion of the rationale behind using a weighted average of sample strata and grand means to estimate population strata means appears in the section on estimation of totals in future random samples.

In the table which follows, a measure of the accuracy of each type of estimator of totals for each of the four variables is listed. In Table 8.3, for a given type of estimator and variable, the number listed under strata j is

$$1/2 \sum_{i=1}^{2} |N_j y_{ij} - N_j \overline{x}_j|,$$

the average absolute deviation over the two random samples of the estimate $N_j y_{ij}$ of the jth strata total from the actual jth strata total $N_j \overline{x}_j$ of the respondent population. The number in parentheses at the bottom of each column in Table 8.3 is the actual respondent population strata total $N_j \overline{x}_j$ for each strata for the given variable. Comparing the average deviation of an estimate with the corresponding actual total gives an indication of the level of accuracy of the type of estimator. All numbers in the table are rounded to two significant

figures. The column titled All Strata in Table 8.3 gives the average absolute deviation summed across the five strata. For instance, 190,000 is listed under Damage, Type 1. This is the sum (rounded off) of the five numbers in the row for Damage, Type 1 in Table 8.3. In parentheses at the bottom of each column is the rounded total for that variable in the respondent population (totaled across strata).

As can be seen from the last column of Table 8.3, Type 1 estimators are best for the Damage and Cost variables in Muskegon County. Type 6 estimators are best for Blufflost and Beachlost totals. That is, these estimators were best when the results of two 60% random samples of the respondent population were averaged. If we assume that the respondent population is a representative 60% sample of the entire population of lakeshore property owners in Muskegon County for these four variables, then we may use these results to estimate county totals.

According to our information, the numbers of lakeshore property owners in Muskegon County in each reach are as listed in Table 8.4.

Means of the four variables for the respondent population are given in Table 8.5.

TABLE 8.3 Absolute Deviations of Totals Averaged Over Two 60% Stratified Random Samples of Respondents in Muskegon County

	Type	Strata (Reach Number) 1	2	3	4	5	All Strata
Damage	1	94,000	36,000	20,000	6,000	33,000	190,000
	2	130,000	11,000	8,000	68,000	104,000	320,000
	3	110,000	16,000	31,000	37,000	62,000	260,000
	4	120,000	12,000	20,000	53,000	83,000	290,000
	5	91,000	25,000	43,000	22,000	42,000	220,000
	6	120,000	12,000	17,000	62,000	130,000	340,000
	7	120,000	7,000	17,000	56,000	87,000	290,000
Respondent Total		(340,000)	(540,000)	(270,000)	(51,000)	(130,000)	(1,300,000)
Cost	1	97,000	92,000	29,000	8,000	48,000	270,000
	2	160,000	180,000	84,000	110,000	140,000	670,000
	3	130,000	110,000	76,000	55,000	92,000	460,000
	4	150,000	140,000	80,000	83,000	110,000	560,000
	5	110,000	73,000	73,000	27,000	70,000	350,000
	6	150,000	120,000	81,000	100,000	150,000	600,000
	7	150,000	150,000	81,000	88,000	120,000	590,000
Respondent Total		(470,000)	(980,000)	(330,000)	(68,000)	220,000	2,100,000
Blufflost	1	97	450	7	120	150	820
	2	86	430	300	110	150	1,100
	3	91	330	150	66	150	790
	4	89	290	220	58	150	810
	5	94	390	70	75	150	780
	6	87	310	240	53	89	780
	7	88	320	240	56	150	850
Respondent Total		(950)	(3,000)	(1,000)	(500)	(950)	(6,400)
Beachlost	1	190	500	930	590	1,900	4,100
	2	650	590	690	210	1,000	3,100
	3	350	280	810	230	1,100	2,800
	4	500	420	750	93	790	2,600
	5	200	390	870	410	1,500	3,400
	6	560	330	740	170	350	2,200
	7	530	450	740	120	830	2,700
Respondent Total		(2,300)	(7,700)	(4,700)	(1,500)	(4,400)	(21,000)

TABLE 8.4 Numbers in Each Strata, Muskegon County

Strata (Reach)	Respondents	Nonrespondents	Total
1	43	25	68
2	112	76	188
3	58	44	102
4	25	12	37
5	49	58	107
All Strata	287	215	502

TABLE 8.5 Means for Respondents in Muskegon County

Strata	Damage	Cost	Bluffloss	Beachloss
1	7800	11000	22	54
2	4800	8800	26	69
3	4600	5700	17	81
4	2000	2700	20	61
5	2600	4400	19	90
Grand Mean	4700	7300	22	70

Treating the respondent population now as the 60% sample, we use the means in Tables 8.5 and the best type of estimator from Table 8.3 to obtain estimates of total losses for Muskegon County. We use the Type 1 estimator for the Damage and Cost variables and the Type 6 estimator for the Blufflost and Beachlost variables. Note that the population for which we are obtaining estimates includes all of Muskegon County's lakeshore property owners and so from Table 8.4 we have $N_1 = 68$, $N_2 = 188$, $N_3 = 102$, $N_4 = 37$, and $N_5 = 107$ as the numbers in the strata to be used in these estimators and $N = 502$.

The estimates of total damages obtained by using Method 2 of estimation appear in Table 8.6.

TABLE 8.6 Estimated Total Damages for Muskegon County (Method 2)

Strata	Damage	Cost	Bluffloss	Beachloss
1	530,000	750,000	1,500	4,600
2	900,000	1,700,000	4,200	13,000
3	470,000	580,000	2,100	7,300
4	74,000	100,000	780	2,600
5	280,000	470,000	2,200	7,900
All Strata	2,300,000	3,600,000	11,000	35,000

This second method for estimating total damages may be applied to the other five counties in an analogous manner. For each county we "simulate" response to the mailed questionnaire with two stratified (by reach) random samples of size given by the response rate in that county. For example, since 56% of the population responded in Manistee County, two 56% stratified random samples were taken from its respondent population. Tables similiar to Table 8.3 could be constructed for each of the remaining counties. However, it is only the last column in Table 8.3, the absolute deviations summed over all strata, that is important for choosing the estimator type for a variable. Table 8.7 summarizes this information for the five remaining counties. For each county in Table 8.7, the column below it should be interpreted the same way as the "All Strata" column of Table 8.3 for Muskegon County. The number in parentheses below each county name represents the questionnaire response rate or the size of the stratified random samples taken from the respondents in that county. Again respondent population totals are given below in parentheses.

TABLE 8.7 Absolute Deviations of Totals Summed Over Strata

(Figures are averages from two stratified random samples)

	Type	Alcona (.6)	County Chippewa (.62)	Huron (.75)	Manistee (.56)	Schoolcraft (.34)
<u>Damage</u>	1	58,100	197,000	56,000	157,000	26,000
	2	68,000	359,000	100,000	324,000	32,000
	3	55,300	252,000	53,000	116,000	24,700
	4	61,500	304,000	77,000	220,000	26,000
	5	55,000	217,000	44,000	81,600	24,700
	6	60,000	288,000	81,000	253,000	26,700
	7	62,904	314,000	81,000	241,000	27,400
	8	71,100	302,000	81,000	296,000	27,400
Respondent Total		(382,000)	(1,100,000)	(404,000)	(788,000)	(41,000)
<u>Cost</u>	1	81,400	256,000	82,000	288,000	24,900
	2	134,000	690,000	246,000	431,000	34,000
	3	80,000	440,000	121,000	151,000	23,700
	4	104,000	564,000	179,000	282,000	26,900
	5	72,158	331,000	90,000	128,000	23,400
	6	106,000	496,000	184,000	331,000	26,100
	7	110,000	589,000	193,000	311,000	28,300
	8	99,200	593,000	193,000	388,000	28,300
Respondent Total		(539,000)	(1,580,000)	(646,000)	(1,140,000)	(47,000)
<u>Bluffloss</u>	1	327	292	340	436	129
	2	313	1,460	820	1,210	213
	3	318	595	440	699	133
	4	315	872	630	934	156
	5	322	336	350	499	120
	6	323	655	640	1,070	141
	7	316	1,580	660	990	167
	8	275	980	660	1,030	167
Respondent Total		(3,500)	(3,600)	(2,850)	(3,200)	(500)
<u>Beachloss</u>	1	625	774	680	1,380	898
	2	811	3,500	2,190	1,780	646
	3	658	2,060	960	1,220	715
	4	729	2,780	1,560	1,400	663
	5	642	1,350	610	1,210	780
	6	763	2,300	1,640	1,600	687
	7	745	2,920	1,680	1,480	651
	8	812	2,950	1,680	1,450	651
Respondent Total		(11,600)	(8,300)	(6,080)	(6,800)	(1,800)

As can be seen from Table 8.7, the "best" type of estimator varies from county to county for each of the variables. For example, in Alcona County estimator types 5, 5, 8, and 1 best predict the four variables. If different estimator types are used to predict Damage and Cost in a particular county, then we must make sure that Damage estimates do not exceed Cost estimates. Such was the case in stratum 5 of Huron County where Damage was estimated as 113,000 and Cost as 109,000. In such a situation the most sensible correction is to inflate the Cost estimate to 113,000. Table 8.8 below summarizes the "best" type of estimator for each variable and county.

TABLE 8.8 "Best" Estimator Types for Each County

	Alcona	Chippewa	Huron	Manistee	Muskegon	Schoolcraft
Damage	5	1	5	5	1	5
Cost	5	1	1	5	1	5
Bluffloss	8	1	1	1	6	5
Beachloss	1	1	1	5	6	2

The "best" estimator types presented in Table 8.8 can now be applied to the respondent population data to predict strata-wide total estimates for each county just as we did for Muskegon County (Table 8.9).

TABLE 8.9 Estimated Total Damages

County	Strata	Damage	Cost	Bluffloss	Beachloss
Alcona	1	521,000	691,000	4,500	16,000
	2	67,000	90,000	760	2,400
	3	56,000	156,000	510	1,000
	Total	644,000	937,000	5,800	19,000
Chippewa	1	22,500	22,500	0	0
	2	54,000	61,000	110	130
	3	966,000	1,603,000	3,300	8,600
	4	59,000	70,000	420	710
	5	0	0	0	0
	6	553,000	687,000	1,700	3,600
	7	102,000	105,000	120	150
	Total	1,758,000	2,549,000	5,700	13,200
Huron	1	1,060,000	2,260,000	5,100	16,000
	2	454,000	836,000	5,800	8,000
	3	740,000	970,000	7,100	14,000
	4	1,090,000	1,300,000	5,100	6,000
	5	113,000	113,000*	1,900	8,000
	Total	3,450,000	5,480,000	25,000	53,000
Manistee	1	247,000	65,000	1,260	3,040
	2	132,000	147,000	960	1,100
	3	26,000	33,000	130	270
	4	61,000	113,000	610	1,230
	5	321,000	461,000	1,560	2,330
	6	49,000	71,000	50	370
	7	438,000	561,000	510	2,200
	8	194,000	369,000	780	1,740
	Total	1,468,000	2,118,000	5,850	12,300
Schoolcraft	1	67,000	75,000	700	1,700
	2	40,000	51,000	540	2,580
	3	3,000	3,000	30	260
	4	4,000	5,000	140	650
	5	700	800	10	120
	Total	115,000	135,000	1,400	5,300

*Inflated from 109,000 to equal Damage estimate.

Finally, let us consider a third method applicable to only Muskegon County. This method makes use of information obtained from the answers given by fourteen nonrespondents during personal interviews. These fourteen nonrespondents represent a random sample of all nonrespondents, with at least two from each reach. In using

responses to personal interviews to estimate what nonrespondents would have answered if they had completed the mailed questionnaire, we must bear in mind that there may be differences in the way people respond to mailed questionnaires and personal interviews. We have a guide to these differences in the comparison of answers by the thirty-four respondents who were also given a personal interview. In Table 8.10 are listed the means of the four variables for these thirty-four respondents from both the questionnaire and the interview. Also listed is the ratio of the questionnaire mean to the interview mean, rounded off.

TABLE 8.10 Means for the Thirty-four Respondents Who Were Also Interviewed

Variable	Questionnaire Mean	Interview Mean	Ratio
Damage	4974	4466	1.1
Cost	7006	7962	0.9
Blufflost	21	25	0.8
Beachlost	48	48	1.0

Note that in the interview setting Damage represents the sum of responses to questions 19 and 21 of the personal interview form. Cost represents Damage plus the dollar amounts answered for question 22. Blufflost is the answer to question 40 and Beachlost the answer to question 38 of the personal interview form.

Nonrespondent data from the personal interview was adjusted as follows to predict what their responses would have been had they returned the questionnaire. Interview damage values were multiplied by 1.1, Cost values by 0.9, Blufflost values by 0.8, and Beachlost values remain unchanged.

The resulting nonrespondent grand means for the adjusted responses appear in Table 8.11, rounded off to the nearest whole number.

TABLE 8.11 Nonrespondent Adjusted Grand Means

Damage	Cost	Blufflost	Beachlost
3216	4836	35	62

To estimate strata-wide nonrespondent damages for all of Muskegon County these means are multiplied by the number of nonrespondents in each strata. These means are multiplied by 25 to obtain an estimate for strata 1, by 76 for strata 2, by 44 for strata 3, by 12 for strata 4, and by 58 for strata 5. County-wide means rather than strata means were used since there were so few interviews of nonrespondents in each reach. These estimated totals appear in Table 8.12.

TABLE 8.12 Estimates of Totals for All Nonrespondents in Muskegon County

Strata	Damage	Cost	Blufflost	Beachlost
1	80,000	120,000	800	1,600
2	240,000	370,000	2,700	4,700
3	140,000	210,000	1,500	2,700
4	39,000	58,000	420	740
5	190,000	280,000	2,000	3,600
Totals	690,000	1,000,000	7,500	13,000

Suppose values in Table 8.12 really do provide reasonable estimates of the total losses which would have been listed by nonrespondents on the mailed questionnaire. A comparison of these values with those given by respondents listed under Muskegon County in Table 8.1 indicates that the total losses for the respondents is 60% of the total losses for respondents plus Table 8.12 estimates of total losses for nonrespondents. That is, for the damage variable, 1,322,000 (from Table 8.1) is approximately 60% of (1,322,000 + 690,000). This suggests that if Table 8.12 provides reasonable estimates of the responses of nonrespondents, then the respondents really are a 60% representative sample of the entire county population--exactly the assumption

used for Method 2. However, we have no way of knowing whether the values in Table 8.12 adequately represent the responses that nonrespondents would have given on the questionnaire.

This final note concerns the accuracy of estimates from the second and third extrapolation methods. Errors in the prelist of population totals could severely bias these estimates, so considerable care should be taken to assure the accuracy of the prelist.

9.0 SAMPLING PLAN FOR FUTURE SURVEYS

The population of interest in the survey is shoreline property owners within reaches of a given county.

1. Planning of survey

Our recommended sample design is the following: Divide the shoreline in a given county into several disjoint reaches. Make a list of all the shoreline property owners in the county. The list should include for each owner name, address, ID number and reach number.

2. Method of sampling selection

For this project we use stratified sampling, with the lakeshore property owners stratified by reach. One of the two sampling schemes I and II described below should be selected as the sampling scheme for the county. Then that scheme is used to obtain a random sample from each of the reaches of the county.

Sampling Scheme I: Simple random sampling within each reach

The steps for obtaining a simple random sample for each reach of the county are as follows:

1) Select one of the reaches of the county.
2) From the number of lakeshore property owners in this reach decide what the size of the random sample from the reach should be. (See part 3, "Size of sample," for recommended sample size.)
3) Use method a or b below to select a random sample of lakeshore property owners in the reach.
4) Repeat steps 1-3 above for each reach in the county.

The following are two methods for obtaining a simple random sample within a reach:

a) Use of a random number table

Suppose we are selecting a random sample from reach i. Let N_i be

the total number of lakeshore property owners in reach i. Suppose it is decided in step (2) above that the random sample from reach i is to have size n_i. Number the property owners in reach i from 1 to N_i. Suppose that N_i is a number with k digits. Obtain a random number table, such as the Rand Number Table by Rand Corporation. Find the first numbers in the table and look at the first k digits. If this k-digit number is N_i or less, select the property owner in reach i corresponding to this number. If the number is greater than N_i, skip it. Go on to the next k digits. If this k-digit number is the same as the number already seen, skip it. If this number is less than or equal to N_i, select the property owner in reach i corresponding to this number. If the number is greater than N_i, skip it. Go on to the next k digits. Repeat this process until a random sample of size n_i has been selected from reach i.

As an example, suppose that N_i is 500 and we want a sample of size $n_i = 3$. The following is a section of a random number table:

```
13 70 43 . . . .
26 99 82 . . . .
72 53 95 . . . .
22 08 08 . . . .
21 61 90 . . . .
. . . . . . . . .
```

The units in the reach with numbers 137, 269, and 220 are selected (725 is excluded because it is larger than 500).

b) Use of Midas

We may use a Midas command to obtain random samples of the appropriate sizes from each reach in the county. Suppose variable 80, V80, equals i if a property owner belongs to reach i. If the following Midas command is given:

CODE BYSTRATA V100=RANDOM SIZES=3 STRATA=V80

the computer will select a random sample of size 3 from each reach and assign V100 the value 1 if a case is in the sample and will code the case as missing otherwise. The CODE command can also be used to select

certain percentages from each reach for the random sample. See the Midas manual for details regarding use of the CODE command for selecting random sample.

Sampling Scheme II: Systematic sampling within each reach

If we would like our random sample for each reach to be evenly distributed with respect to some characteristic, then systematic sampling is ideal. For example, suppose we would like our sample to be evenly distributed along the shoreline within each reach. Then the steps for selecting a stratified random sample for each reach are as follows:

1) Select one of the reaches of the county, say reach i.

2) From the number N_i of lakeshore property owners in reach i, decide what the size n_i of the random sample from this reach should be.

3) Suppose $N_i = 500$ and $n_i = 10$. Select a random number between 1 and 50. If 32 is the random number, the sample consists of 32, 82, 132, 182, ..., 482. For general N_i and n_i, select a random number k between 1 and m_i, where $m_i = [N_i/n_i]$ represents the greatest integer less than or equal to N_i/n_i (which may not be a whole number). Then the random sample consists of $k, k + m_i, k + 2m_i, \ldots, k + (n_i - 1)m_i$.

4) Repeat steps 1-3 above for each reach in the county.

3. Size of sample

From the pilot study, we recommend that the size of the sample in each reach be 20% of the population size in the corresponding reach or 30, <u>whichever is greater</u>. However, if this "sample size" is greater than the reach size, sample the entire reach. In other words

Sample size = minimum (x, reach size) where

x = maximum (20% of reach size, 30).

Example: N = reach size   s = sample size

(1) N = 400

20% of N = 80

x = max (80, 30) = 80

$$s = \min(80, 400) = 80$$

(2) $N = 100$

20% of $N = 20$

$$x = \max(20, 30) = 30$$

$$s = \min(30, 100) = 30$$

(3) $N = 20$

20% of $N = 4$

$$x = \max(4, 30) = 30$$

$$s = \min(30, 20) = 20$$

Although samples of size 30 within each reach are believed to be sufficient, a 20% sample figure is included as a safety factor for reaches of size greater than 150. For densely populated counties, responses may differ appreciably from those observed in Muskegon County, for instance, and a larger sample may be needed to determine the extent of losses due to high lake levels.

4. On obtaining the entire random sample

The reader by now will have noticed that in many parts of this analysis, interpretations were confused and conclusions impossible to draw because of the problem of nonrespondents. When a random sample is drawn and questionnaires mailed or interviewers sent out, there are bound to be members of the sample who for one reason or another do not wish to respond. Just forgeting these nonrespondents and treating the remainder of the sample as the random sample can lead to very biased results if nonrespondents represent a portion of the population which varies significantly from respondents with respect to characteristics of interest. For instance, suppose most nonrespondents did not respond because they suffered no damages at all. Then the average damages reported by respondents will probably be much higher than the average damages suffered by all lakeshore property owners in the county. On the other hand, suppose most nonrespondents did not respond because their properties had been damaged beyond hope by the high lake levels and they did not feel it was worthwhile to respond. Then the average of damages suffered by respondents will

probably be much lower than the average damages suffered by all lakeshore property owners in the county.

Now suppose that a member of the random sample does not respond and so you send a questionnaire to a different member of his reach, or interview the property owner next door. Such practices destroy the randomness of the sample and cause the same problems of analysis described in the previous paragraph.

Therefore, for proper interpretation of the results of a random sample, responses must be obtained from all members of the sample selected. This may entail mailed reminders, making telephone calls, or calling on the nonrespondent at home. Whatever the means, responses must be obtained from all members of the random sample selected, to allow for any kind of meaningful analysis of the data obtained.

5. Further recommendations

A census of "outliers" not found to be coding errors should be conducted. (See section 5 for a discussion of outliers.)

We recommend that a random sample of 40 of the original random sample be selected for personal interviews, with the 40 distributed throughout all reaches of the county. These personal interviews will serve as a check on responses to a mailed questionnaire (assuming that the mailed questionnaire was the initial method used to obtain responses). Recall the recommendations in section 2 that questions be asked in identical form in interview and questionnaire settings so that any differences in responses may be attributed to differences in wording of questions. The same remarks as in part 4 above apply with regard to interviewing the entire sample selected.

In view of the present high water levels in the Great Lakes, we propose that our sampling plan be used in as many as possible (preferably all) of the remaining counties bordering on the Great Lakes. Cross-county comparisons would not be possible otherwise. Also, we expect that response rates to questionnaires would drop appreciably with water levels if the sampling process were to continue over a

period of years. Therefore, to maximize information about damages being caused by current high water levels, we recommend that all sampling be completed as soon as possible.

10.0 ESTIMATES OF TOTALS FOR FUTURE SURVEYS

In this section we wish to recommend appropriate estimates of total damages suffered in a county, for use when future surveys are conducted. All recommendations are based on the assumption that when a random sample has been chosen from a county, responses are obtained from <u>all</u> members of the random sample so that there are no problems with nonrespondents. (See section 9 on sampling for a discussion of the problem of nonrespondents.)

Our search for appropriate estimates of total damages suffered within a county was based upon responses to the mailed questionnaire in three counties: Muskegon, Manistee, and Alcona. The four variables included in this discussion will be called Damage, Cost, Blufflost, and Beachlost.

Damage is the sum of all losses due to flood damage and erosion damage as listed by the respondents in question B2 of the mailed questionnaire. Cost is the sum of Damage plus the cost of labor and materials for protective actions, as listed in B4 of the questionnaire. Blufflost and Beachlost are the answers to D3 and D4 of the questionnaire, respectively. See Appendix V-c for a copy of the mailed questionnaire.

Our goal was to find the best methods of estimating strata-wide totals of these four variables, where the strata correspond to the reach number classifications of the county. The procedure used for each of the three counties listed above was as follows.

Five random samples of the population of respondents were drawn, such that each random sample consisted of approximately 20% of the population in each strata of the county. That is, the first random sample was obtained by randomly drawing 20% of the members of the stratum with reach number 1, 20% of the members of the stratum with reach number 2, and so on for each strata of the county. This process was repeated until five random samples were obtained.

Before explaining the methods tried for estimating totals, we need some notation. Suppose there are k reaches in the county. For a given variable, let $\hat{x}_{ij}$ stand for the mean of that variable in the

jth stratum for the ith random sample. Suppose we are considering the variable Damage. Then, for instance, $\hat{x}_{11}$ stands for the mean of the Damage values in the first stratum (reach number 1) of the first random sample; $\hat{x}_{12}$ stands for the mean of the Damage values in the second stratum (reach number 2) of the first random sample; $\hat{x}_{1k}$ stands for the mean of the Damage values in the kth stratum (reach number k) of the first random sample. Similarly for the other random samples.

Again, for a given variable, let $\hat{x}_i$ stand for the grand mean for that variable in the ith random sample. If we are considering the variable Damage, then $\hat{x}_3$, for instance, stands for the grand mean of the Damage values for all respondents in the third random sample. Thus, the strata are ignored and the entire random sample used to compute the grand means $\hat{x}_i$.

Finally, for a given variable, let $\bar{x}_j$ stand for the mean for that variable within the jth stratum of the original population. The $\bar{x}_j$'s are the actual stratum means of the respondent population.

Let N be the total number of respondents for the county under consideration. Let N_i be the number of respondents from reach number i, $i = 1, \ldots, k$. Then $N = N_1 + N_2 + \ldots + N_k$. Note that when using a real random sample from a future county, N will be the total number of lakeshore property owners in the county and N_i will be the total number of lakeshore property owners in reach i for the county.

We are trying to estimate total losses within each strata of the population of respondents. Therefore, we are going to try to estimate $N_1\bar{x}_1, N_2\bar{x}_2, \ldots, N_k\bar{x}_k$ for each of the four variables.

Now let y_{ij} be the estimate of $\bar{x}_j$ for the ith sample. Then $N_j y_{ij}$ is used to estimate $N_j\bar{x}_j$. For example, y_{23} is the estimate of the mean $\bar{x}_3$ of the third strata of the population, obtained from the second random sample. $N_3 y_{23}$ is the second random sample estimate of the third strata population total $N_3\bar{x}_3$.

There were eight methods of estimation tried for each of the four variables. These methods are tabulated in Table 10.1 below.

TABLE 10.1

Method	Estimate of $\bar{x}_j$
1	$y_{ij} = \hat{x}_{ij}$
2	$y_{ij} = \hat{x}_i$
3	$y_{ij} = (1/2)\ \hat{x}_{ij} + (1/2)\ \hat{x}_i$
4	$y_{ij} = (1/4)\ \hat{x}_{ij} + (3/4)\ \hat{x}_i$
5	$y_{ij} = (3/4)\ \hat{x}_{ij} + (1/4)\ \hat{x}_i$
6	$y_{ij} = (N_j/N)\ \hat{x}_{ij} + (1 - N_j/N)\ \hat{x}_i$
7	$y_{ij} = (1/5)\ \hat{x}_{ij} + (4/5)\ \hat{x}_i$
8	$y_{ij} = (1/k)\ \hat{x}_{ij} + (1 - 1/k)\ \hat{x}_i$

Method 1 estimates population strata means with sample strata means. Method 2 ignores strata differences and uses sample grand means to estimate population strata means.

Methods 3-7 each use some sort of average of sample strata and grand means to try and achieve better estimates of population strata means than are obtained in Methods 1 and 2. Since some sort of average of $\hat{x}_{ij}$ and $\hat{x}_i$ uses more information than using either alone, it was expected that methods 3-7 would provide better estimates than methods 1 and 2. This proved in general to be the case, as will be seen later.

Method 3 assigns equal weights of 1/2 to each of $\hat{x}_{ij}$ and $\hat{x}_i$. Methods 4, 5, and 7 assign weights of 1/4, 3/4; 3/4, 1/4; and 1/5, 4/5, respectively, to $\hat{x}_{ij}$ and $\hat{x}_i$. Method 6 weights $\hat{x}_{ij}$ according to the relative size of strata j in the respondent population. Method 8 assigns weight 1/k to $\hat{x}_{ij}$ and weight 1 - 1/k to $\hat{x}_i$.

As an example, Muskegon County has k = 5 reaches. There were $N_3 = 58$ respondents from reach number 3 and N = 287 respondents for the entire county. Then for method 6, $y_{13} = (58/287)\hat{x}_{13} + (1 - 58/287)\ \hat{x}_1$, for Muskegon County. For method 7, $y_{13} = (1/5)\ \hat{x}_{13} + (4/5)\ \hat{x}_1$. Note that for Muskegon County and other counties with five reaches (k = 5), methods 7 and 8 are equivalent.

Why would we want to use a weighted average of sample strata and grand means to estimate population strata means? A pragmatic answer, and the only meaningful one in any practical sense, is that by incor-

porating the additional information provided in the grand means into the estimate, better estimates are obtained. Evidence of this improvement is shown in Tables 10.3-10.6. Theoretical results have demonstrated that a weighted average of strata and grand means in general draws the estimates closer to the true population strata means than the sample strata means alone.

Another reason, besides the theoretical result, for wishing to use weighted averages of sample strata and grand means to estimate population strata means, is the result of the analyses of variance we performed on the natural logarithms of these four variables. Suppose that variables 501-504 represent the natural logarithms of the variables Damage, Cost, Blufflost, and Beachlost, respectively. An analysis of variance will be performed in Midas for each of these four variables in response to the following command:

ANOVA VAR=501-504 STRATA=V80

The stratifying variable is variable 80, reach number classification. This allows the testing of whether there is a significant difference between the strata means for each of variables 501-504.

Note that we performed the analysis of variance on the logarithm of each variable. This is because these four variables each have distributions well approximated by a lognormal distribution (see section 3). Therefore, the logarithm of the variables are approximately normally distributed. Since the analysis of variance is based on the assumption that the variables are normally distributed, the logarithm of the variables is used in the analysis of variance. This is an important reason for discovering and noting that these variables do have approximate lognormal distributions.

Table 10.2 shows the results of the analysis of variance for variables 501-504 for Alcona, Manistee, and Muskegon Counties.

TABLE 10.2 ANOVA

County	Variable	Significance	Eta-square
Alcona	Log of Damage	.3818	.0142
	Log of Cost	.1829	.0232
	Log of Blufflost	.1196	.0326
	Log of Beachlost	.2941	.0108
Manistee	Log of Damage	.4059	.0580
	Log of Cost	.3241	.0614
	Log of Blufflost	.0251	.1543
	Log of Beachlost	.6304	.0499
Muskegon	Log of Damage	.0241	.0595
	Log of Cost	.0503	.0460
	Log of Blufflost	.0464	.0512
	Log of Beachlost	.0882	.0405

The column titled Significance in Table 10.2 gives the probability that differences between the strata means equal to or greater than the those observed could have occurred by chance. Therefore, if the significance level is very small, then there is reason to conclude that there are significant differences between the means of the variable for different strata.

The numbers in the Eta-square column of Table 10.2 provide a measure of the proportion of the total variation (between all responses for a variable) which is explained by the variation in responses between reaches (as compared to the variation in responses for respondents within a reach). A small value of eta-square indicates that within reach fluctuation of responses accounts for as much as or more of the variation between all responses than does between reach fluctuation. That is, while there is a lot of variation in responses within a reach, the reaches may not be all that different from each other with regard to responses.

Although some of the significance levels in column 2 of Table 10.2 are small, only three are less than .05 and none are less than .01. Also, the eta-square values are all very small. Therefore, the results of the analyses of variance for all three counties indicate that there is information to be obtained by using the sample grand means (and thus all the sample data) in estimates of population strata means.

In Tables 10.3-10.6 which follow, a measure of the accuracy of each method of estimation of totals for each of the four variables is listed for each of the three counties. In Tables 10.3-10.5, for a given method and variable, the number listed under strata j is

$$1/5 \sum_{i=1}^{5} |N_j y_{ij} - N_j \bar{x}_j|,$$

the average absolute deviation over the five random samples of the estimate $N_j y_{ij}$ of the jth strata total from the actual jth strata total $N_j \bar{x}_j$. The number in parentheses at the bottom of each column in Tables 10.3-10.5 is the actual population strata total $N_j \bar{x}_j$, for each strata, for the given variable. Comparing the average deviation of an estimate with the corresponding actual total gives an indication of the level of accuracy of the method of estimation. All numbers in the tables are rounded to two significant figures.

Table 10.6 gives the average absolute deviation summed across the five strata. For instance, 370000 is listed under Damage, Method 1 for Muskegon County. This is the sum (rounded off) of the five numbers in the row for Damage, Method 1 in Table 10.3. In parentheses at the bottom of each column in Table 10.6 is the rounded total for that variable in the population (totaled across strata).

As can be seen from Table 10.6, different methods gave best estimates of totals for different variables in different counties. Choosing the methods which produced the smallest absolute deviations summed across strata (as listed in Table 10.6), we see that in Alcona County, methods 4 and 7 performed best for Damage, method 3 for Cost, method 8 for Blufflost and method 3 for Beachlost. In Manistee County, best estimates were given by method 8 for Damage and Cost and by method 3 for Blufflost and Beachlost. In Muskegon County, method 3 worked the best for Damage and Cost while method 6 worked best for Blufflost and Beachlost. On the basis of these results, we recommend that when random samples are taken for future counties, 20% simulations be performed as described in this section and best methods selected (on the basis

of these simulations) for estimating Damage, Cost, Blufflost, and Beachlost. Estimates of total losses suffered in the county may then be obtained in the manner described in section 8.

A final note is justified concerning the use of 20% random samples in this analysis. In Muskegon County, for instance, where there were only 25 respondents in reach number 4, samples of size smaller than 20% contained fewer than four respondents. With so small a sample, estimates of totals are very inaccurate. If a strata in the population of interest contains very few persons, even a 20% sample may be inadequate. In such a case, including the entire strata or taking a random sample which includes more than 20% of the strata would be called for. The idea is to make the sample sizes as small as possible (to reduce sampling costs) while still maintaining acceptable accuracy in estimation. These ideas form the basis for the recommended sampling sizes provided in section 9.

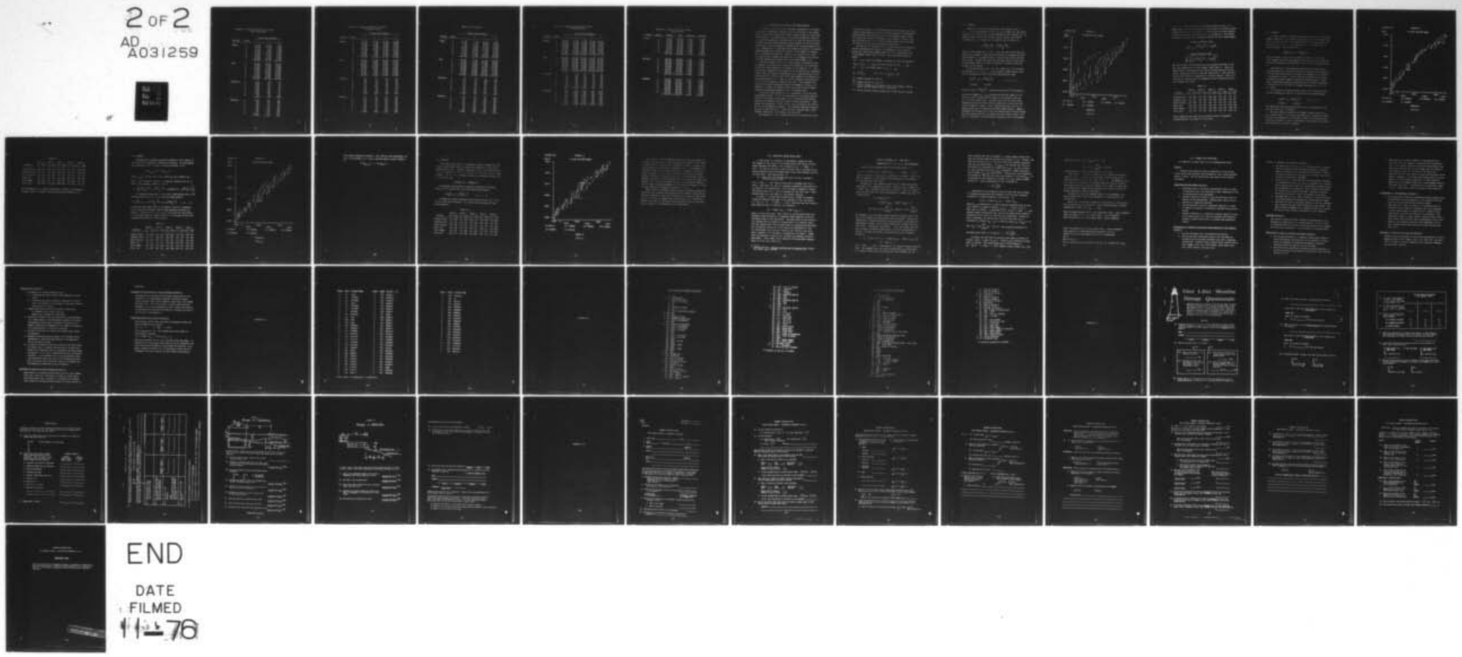
Section 9 on sampling plan discusses the recommended sample size for future surveys. This plan will always result in strata samples larger than those obtained in the 20% random samples used in simulations to evaluate schemes for estimating totals. Therefore, the methods of estimation judged as good in 20% simulations should do <u>better</u> in practice since the sample sizes will be larger than those used in the simulations and so better estimates of population totals can be obtained. The absolute deviations presented in Tables 10.3-10.6, therefore, are conservative estimates of the error to be expected in actual estimation.

In this connection, notice that the absolute deviations given in Tables 10.3-10.6 are those obtained by simulated sampling from the respondents of the mailed questionnaire for each of the three counties. These values <u>cannot</u> be applied to the counties as a whole since they were obtained from samples from the counties. See section 8 for our proposed estimates of actual total damages for these three counties.

AD-A031 259 MICHIGAN UNIV ANN ARBOR COASTAL ZONE LAB F/G 13/2
PILOT STUDY PROGRAM, GREAT LAKES SHORELAND DAMAGE STUDY. APPEND--ETC(U)
MAY 76 E D ROTHMAN, S BARTOLD, R BUTLER DACW23-75-C-0027

UNCLASSIFIED NL

2 OF 2
AD A031259



END
DATE
FILMED
11-76

TABLE 10.3 Absolute Deviations of Totals
for Alcona County

Variable	Method	Strata (Reach Number) 1	2	3
Damage	1	36,000	20,000	7,000
	2	24,000	14,000	2,000
	3	27,000	4,000	5,000
	4	24,000	5,000	3,000
	5	31,000	11,000	6,000
	6	33,000	10,000	2,000
	7	24,000	7,000	3,000
	8	85,000	12,000	6,000
		(330,000)	(30,000)	(19,000)
Cost	1	78,000	23,000	23,000
	2	57,000	27,000	34,000
	3	68,000	2,000	17,000
	4	62,000	15,000	25,000
	5	73,000	10,000	19,000
	6	75,000	21,000	32,000
	7	61,000	17,000	27,000
	8	114,000	14,000	37,000
		(440,000)	(39,000)	(64,000)
Blufflost	1	1,200	130	90
	2	1,000	160	70
	3	1,100	140	30
	4	1,000	150	40
	5	1,100	140	50
	6	1,100	160	60
	7	1,000	150	40
	8	420	50	40
		(3,000)	(360)	(180)
Beachlost	1	1,000	140	280
	2	580	190	290
	3	790	70	10
	4	680	130	150
	5	900	110	140
	6	930	160	260
	7	660	140	180
	8	1,400	180	140
		(10,000)	(1,200)	(360)

TABLE 10.4 Absolute Deviations of Totals for Mansitee County

Variable	Method	Strata (Reach Number) 1	2	3	4
Damage	1	110,000	56,000	38,000	32,000
	2	94,000	9,000	34,000	48,000
	3	98,000	23,000	31,000	25,000
	4	91,000	7,000	32,000	37,000
	5	100,000	39,000	31,000	25,000
	6	92,000	7,000	33,000	44,000
	7	90,000	4,000	33,000	39,000
	8	65,000	7,000	27,000	37,000
		(140,000)	(35,000)	(14,000)	(33,000)
Cost	1	110,000	55,000	38,000	62,000
	2	100,000	10,000	47,000	37,000
	3	110,000	24,000	38,000	37,000
	4	100,000	10,000	42,000	32,000
	5	110,000	40,000	33,000	49,000
	6	100,000	9,000	46,000	35,000
	7	100,000	8,000	43,000	33,000
	8	96,000	9,000	45,000	34,000
		(200,000)	(36,000)	(14,000)	(70,000)
Blufflost	1	250	240	150	110
	2	190	130	80	120
	3	220	130	90	100
	4	210	80	80	110
	5	230	190	120	80
	6	210	120	80	120
	7	200	90	80	110
	8	220	120	80	110
		(790)	(240)	(100)	(430)
Beachlost	1	410	250	130	210
	2	390	90	150	400
	3	270	120	140	240
	4	330	70	150	320
	5	340	190	130	200
	6	320	80	150	370
	7	340	70	150	330
	8	310	60	170	290
		(1,900)	(290)	(160)	(940)

TABLE 10.4 (continued)

Variable	Method	Strata (Reach Number) 5	6	7	8
Damage	1	100,000	6,000	68,000	73,000
	2	51,000	23,000	88,000	29,000
	3	77,000	14,000	69,000	41,000
	4	64,000	19,000	70,000	25,000
	5	90,000	10,000	69,000	57,000
	6	64,000	22,000	73,000	28,000
	7	62,000	19,000	73,000	24,000
	8	39,000	17,000	98,000	26,000
		(210,000)	(11,000)	(260,000)	(85,000)
Cost	1	130,000	2,000	57,000	130,000
	2	72,000	27,000	100,000	98,000
	3	100,000	14,000	71,000	82,000
	4	86,000	21,000	83,000	85,000
	5	110,000	8,000	64,000	110,000
	6	86,000	26,000	88,000	95,000
	7	83,000	22,000	88,000	87,000
	8	54,000	27,000	96,000	76,000
		(310,000)	(17,000)	(320,000)	(170,000)
Blufflost	1	150	20	140	210
	2	270	110	380	100
	3	140	50	200	150
	4	210	80	290	120
	5	130	20	120	180
	6	200	110	310	100
	7	220	90	310	110
	8	230	110	360	90
		(1,000)	(20)	(270)	(310)
Beachlost	1	220	80	490	330
	2	70	160	60	390
	3	110	40	230	340
	4	60	100	100	370
	5	170	20	360	320
	6	60	150	80	390
	7	70	110	80	370
	8	220	170	220	290
		(1,500)	(80)	(1,100)	(770)

TABLE 10.5 Absolute Deviations of Totals for Muskegon County

Variable	Method	Strata (Reach Number) 1	2	3	4	5
Damage	1	210,000	18,000	95,000	23,000	28,000
	2	140,000	78,000	42,000	64,000	96,000
	3	130,000	36,000	61,000	32,000	48,000
	4	110,000	57,000	50,000	48,000	72,000
	5	170,000	17,000	78,000	23,000	28,000
	6	120,000	45,000	48,000	58,000	80,000
	7	110,000	61,000	48,000	51,000	77,000
		(340,000)	(540,000)	(270,000)	(51,000)	(130,000)
Cost	1	300,000	160,000	110,000	33,000	46,000
	2	160,000	180,000	90,000	120,000	140,000
	3	180,000	160,000	90,000	56,000	68,000
	4	150,000	160,000	90,000	84,000	100,000
	5	240,000	160,000	90,000	32,000	34,000
	6	140,000	160,000	90,000	100,000	110,000
	7	140,000	160,000	90,000	89,000	110,000
		(470,000)	(980,000)	(330,000)	(68,000)	(220,000)
Blufflost	1	220	560	230	330	300
	2	40	450	290	30	150
	3	90	340	170	80	150
	4	40	340	230	40	100
	5	130	450	170	130	200
	6	40	340	230	30	100
	7	40	340	230	40	100
		(940)	(3,000)	(1,000)	(500)	(950)
Beachlost	1	860	1,000	1,700	800	1,500
	2	770	1,100	900	350	1,000
	3	520	1,000	1,300	380	930
	4	560	1,100	1,100	330	880
	5	690	900	1,500	550	1,200
	6	650	1,100	1,100	350	930
	7	600	1,100	1,100	330	930
		(2,300)	(7,700)	(4,700)	(1,500)	(4,400)

TABLE 10.6 Absolute Deviations Summed Across Strata

County	Method	Damage	Cost	Blufflost	Beachlost
Alcona	1	60,000	120,000	1,400	1,400
	2	40,000	120,000	1,200	1,100
	3	40,000	90,000	1,300	900
	4	30,000	100,000	1,200	1,000
	5	50,000	100,000	1,300	1,200
	6	50,000	130,000	1,300	1,400
	7	30,000	110,000	1,200	1,000
	8	100,000	170,000	500	1,700
		(382,000)	(580,000)	(3,500)	(11,600)
Manistee	1	480,000	580,000	1,300	2,100
	2	380,000	490,000	1,400	1,700
	3	380,000	480,000	1,100	1,500
	4	720,000	460,000	1,200	1,500
	5	420,000	520,000	1,100	1,700
	6	360,000	490,000	1,300	1,600
	7	340,000	460,000	1,200	1,500
	8	320,000	440,000	1,300	1,700
		(788,000)	(1,140,000)	(3,200)	(6,800)
Muskegon	1	370,000	650,000	1,600	5,900
	2	420,000	690,000	960	4,100
	3	310,000	550,000	830	4,100
	4	340,000	580,000	750	4,000
	5	320,000	560,000	1,100	4,800
	6	350,000	600,000	740	4,000
	7	350,000	590,000	750	4,100
	8	350,000	590,000	750	4,100
		(1,300,000)	(2,100,000)	(6,400)	(21,000)

11.0 ESTIMATION OF PROPORTIONS FOR FUTURE SURVEYS

In the section on use of the lognormal approximation, we briefly considered estimating for Muskegon County the fraction of the respondents to the mailed questionnaire whose property depth was within some given interval. In this section we extend this estimation procedure to reach-wide estimation of proportions to be used in conjunction with the 20% stratified random sampling scheme proposed for future surveys. Specifically we are interested in answering the following general question. Suppose that a county has been surveyed according to our proposed sampling scheme and a particular variable, total damage for example, fits a lognormal distribution. Then for each reach of the county, how might we best estimate the proportion of the <u>entire</u> reach population that suffers total damage within some fixed range (a, b)?

Our basic objective then is to find a method of estimation that, for example, will accurately estimate the fraction of the population within each reach suffering total damage in the range (a, b). Since the fraction responding in the range (a, b) is the fraction responding below b minus the fraction responding below a, good estimates of proprotions within intervals result from a good method for estimating the stratum population c.d.f. That is, if $F(\cdot)$ estimates the population c.d.f. of total damage for reach #1 then $F(b) - F(a)$ estimates the fraction of reach #1 population whose total damage is within (a, b). Therefore to answer our question above, we need only find the best method for estimating population c.d.f.'s for each reach.

To find this best method of c.d.f. estimation, we use the data from Muskegon County. We let the respondents to the mailed questionnaire be our population and simulate 20% stratified random samples (mimicing the proposed sampling scheme) of these respondents. Then, because the population is known, we are able to evaluate the accuracy of the various methods we propose. The following lognormal variables are used in this evaluation: Property worth, assessed value, Bluff distance, Beach lost, total damage, and total cost.

Four methods of c.d.f. estimation have been considered and their

accuracies evaluated. For each method five 20% stratified (by reach number) random samples were taken from the population of respondents, simulating what will actually occur in our proposed sample survey. For a particular method of estimation, each sample is used to estimate the population c.d.f. of each variable within each strata and the accuracy of each of the estimates is measured.

We restrict our discussion of the various methods of estimation below by estimating the fraction of respondents within reach #1 (strata #1) having total damage within (a, b) or equivalently log (total damage) within (log a, log b). The following notation is needed:

$x_1, \ldots, x_N$ = log (total damage) responses for reach #1 population

$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(N)}$ are ordered values of $\{x_1, \ldots, x_N\}$

$x_1, \ldots, x_n$ = 20% random sample of $\{x_1, \ldots, x_N\}$

$\bar{x}_1 = 1/n \sum_1^n x_j \qquad s_1^2 = 1/(n-1) \sum_{i=1}^n (x_i - \bar{x}_1)^2$

$\bar{x}_1$ = sample average for reach #1

s_1^2 = sample variance for reach #1

$\bar{x}$ = overall average of all values of log (total damage), making use of the 20% samples from all the reaches

s^2 = overal variance estimate making use of data from all reaches

11.1 METHOD 1

Making use of only intra-reach data, the discrete c.d.f. of $x_1, \ldots, x_N$ is approximated by the c.d.f. of a normal $(\bar{x}_1, s_1^2)$ random variable. Our estimate here is the probability that a normal $(\bar{x}_1, s_1^2)$ random variable lies in the interval (log a, log b), which after standardization to a normal (0, 1) variable can be written as

$$\Phi\left(\frac{\log b - \bar{x}_1}{s_1}\right) - \Phi\left(\frac{\log a - \bar{x}_1}{s_1}\right),$$

where Φ is the normal (0, 1) c.d.f. Note that this is the same type estimate as proposed in the section on the use of lognormal distributions; however, it is applied only to reach #1 here rather than the whole county. We will see later that the small amount of data used in this method (20% of 43 cases in reach #1 = 8 cases) limits the accuracy of the method.

To evaluate how well the c.d.f. of a normal $(\bar{x}_1, s_1^2)$ estimates the discrete c.d.f. of $x_1, \ldots, x_N$ we construct Midas scatter plots of the following for each of the five random samples taken:

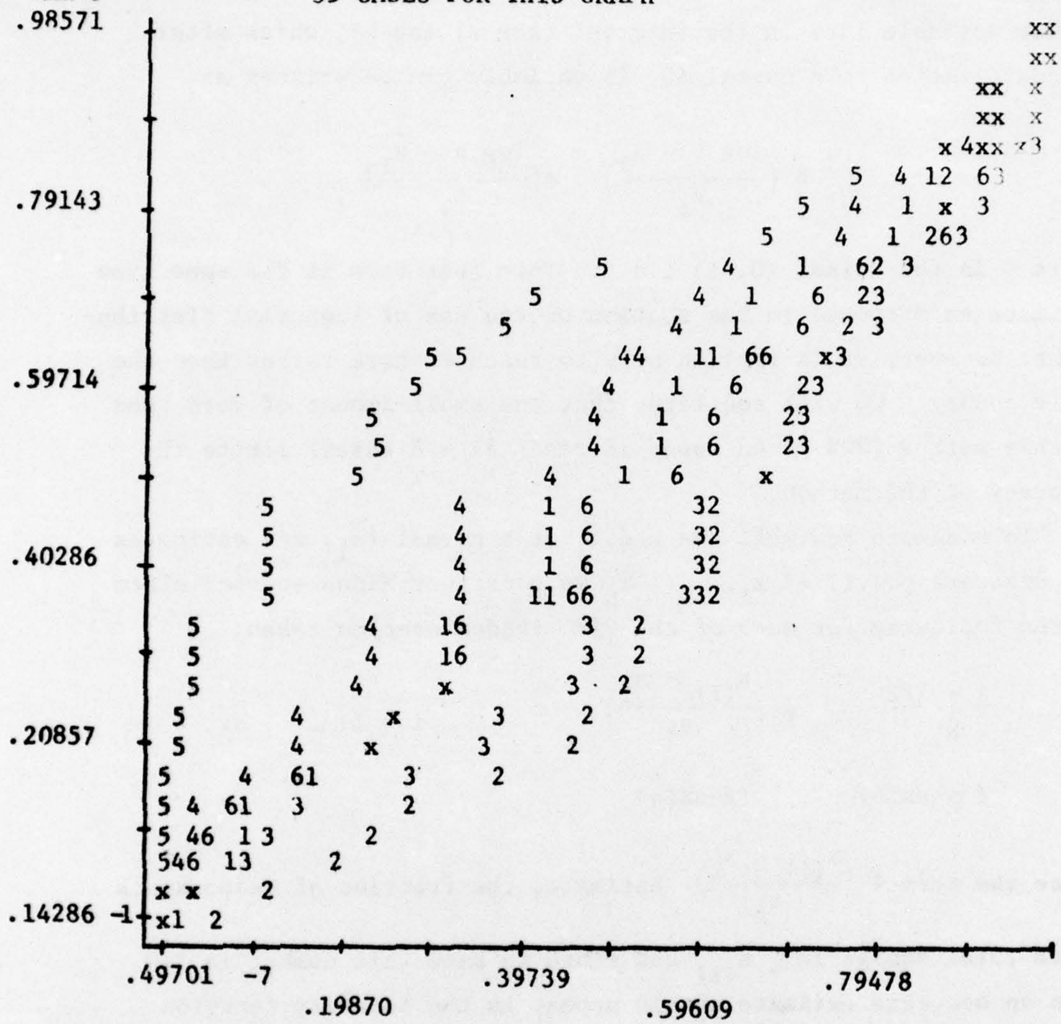
$$\frac{i - 1/2}{N} \quad \text{vs.} \quad \Phi\left(\frac{x_{(i)} - \bar{x}_1}{s_1}\right) \qquad i = 1, \ldots, N.$$

(y-axis) (x-axis)

Since the term $\Phi\left(\frac{x_{(i)} - \bar{x}_1}{s_1}\right)$ estimates the fraction of respondents whose total damage is $\leq x_{(i)}$ and since we know this number is i/N, then an accurate estimate should appear as the identity function (y = x) in the scatter plot. The five plots (one for each random sample) may be plotted simultaneously because the y-axis stays the same from sample to sample. Such a simultaneous plot (Plot 11.1) appears below with 2-6 numbering different random samples. Number 1 represents the average plot of the five samples constructed by averaging the x-values over the five samples for each fixed y-value. Note that "X" represents the simultaneous occurance of two of the numbers

SCATTER PLOT   STRATUM = 1

VAR 6   35 CASES FOR THIS GRAPH



(1) VAR 36   (2) LTOTDAM   (3) LTOTDAM   (4) LTOTDAM

(5) LTOTDAM   (6) LTOTDAM

METHOD #1

PLOT 11.1

1, ..., 6 . The variability of these plots about the line $y = x$ demonstrates the sampling variance of this estimation procedure, which is large in this particular plot. In order to get an overall measure of how close each plot fits the line $y = x$ (a measure of the methods accuracy), we compute the following statistics for each random sample:

Kolmogorov-Smirnoff (K-S)

$$\max_{i=1,\ldots,N} \left| \Phi\left(\frac{x_{(i)} - \bar{x}_1}{s_1}\right) - \frac{i - 1/2}{N} \right|$$

Cramer-Von Mises (C-VM)

$$\sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(\Phi\left(\frac{x_{(i)} - \bar{x}_1}{s_1}\right) - \frac{i - 1/2}{N} \right)^2}$$

The five K-S and five C-VM statistics are averaged resulting in .213 and .136 respectively for total damage, Reach number 1. These two numbers may be interpreted as the overall maximum deviation and the overall expected deviation of the plot respectively. These two statistics may be used for a comparison of the four methods. Table 11.1 provides the average (over five samples) K-S and C-VM statistics for each variable and strata.

TABLE 11.1

	Reach 1		Reach 2		Reach 3		Reach 4		Reach 5	
Variable	K-S	C-VM	K-S	C-VM	K-S	C-VM	K-S	C-VM	K-S	C-VM
Property worth	.154	.086	.174	.085	.208	.120	.249	.152	.192	.135
Assessed value	.171	.119	.097	.060	.192	.098	.356	.185	.191	.101
Bluff distance	.193	.098	.113	.054	.251	.148	.292	.197	.310	.201
Beach lost	.238	.129	.107	.049	.213	.115	.271	.153	.215	.118
Total damage	.213	.136	.180	.095	.125	.066	.278	.154	.232	.154
Total cost	.253	.164	.123	.067	.199	.110	.187	.101	.228	.169

After comparing this table with equivalent tables of subsequent methods, Method 1 was found to be the worst.

11.2 METHOD 2

In direct contrast to Method 1, we make use of all the log (total damage) responses in our random sample from all five reaches and treat them as if they all came from stratum (reach) 1. We approximate the discrete c.d.f. of $x_1, \ldots, x_N$ with the c.d.f. of a Normal $(\bar{x}, s^2)$ random variable. Our estimate here is

$$\Phi\left(\frac{\log b - \bar{x}}{s}\right) - \Phi\left(\frac{\log a - \bar{x}}{s}\right).$$

Such an estimate is the result of fitting a normal curve for log (total damage) using the data sampled from the entire county and predicting the population distributions of each of the five reaches from it.

Justification of this method and the next two methods results from both theoretical considerations and the marginal significance of the F-test from the analysis of variance as explained in the section on estimation of totals. If the F test is not significant then the reaches may represent somewhat artificially created strata. Under these circumstances sampling variability may be reduced considerably by making use of all the data to estimate the population c.d.f. of each reach.

As in Method 1 we evaluate the accuracy of this method by plotting the following for each of the five stratified random samples:

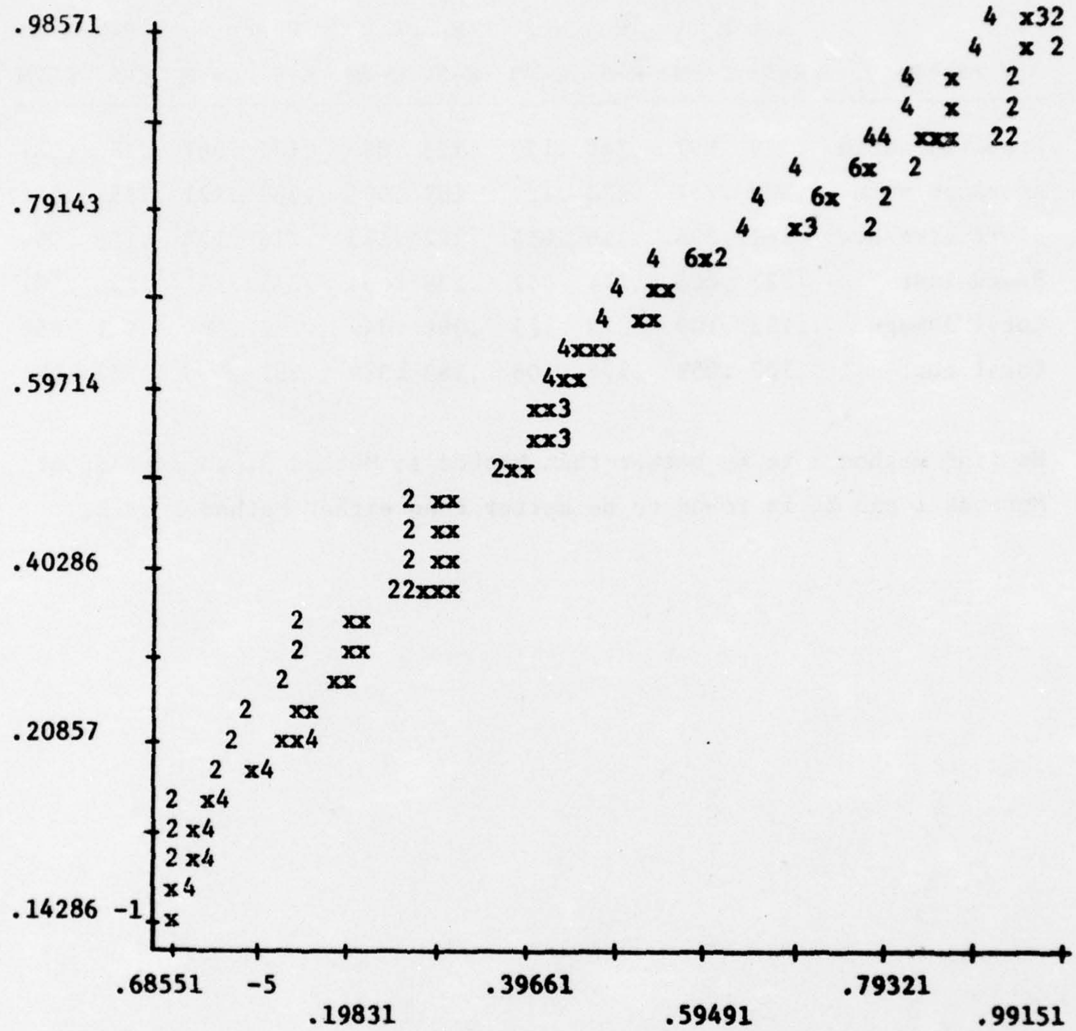
$$\underset{\text{(y-axis)}}{\frac{i - 1/2}{N}} \quad \text{vs.} \quad \underset{\text{(x-axis)}}{\Phi\left(\frac{x_{(i)} - \bar{x}}{s}\right)} \qquad i = 1, \ldots, N$$

The simultaneous plot given below (Plot 11.2) is interpreted in the same way as the plot for Method 1.

This plot shows the method to be slightly biased (the plot is not centered about $y = x$), however the reduction in sampling variability more than compensates for the bias to make this estimator better than that of Method 1. This is reflected in the average K-S and C-VM statistics given in Table 11.2 below:

SCATTER PLOT STRATUM = 1

VAR 6 35 CASES FOR THIS GRAPH



(1) VAR 36 (2) LTOTDAM (3) LTOTDAM (4) LTOTDAM

(5) LTOTDAM (6) LTOTDAM

METHOD #2

PLOT 11.2

TABLE 11.2

Variable	Reach 1 K-S	Reach 1 C-VM	Reach 2 K-S	Reach 2 C-VM	Reach 3 K-S	Reach 3 C-VM	Reach 4 K-S	Reach 4 C-VM	Reach 5 K-S	Reach 5 C-VM
Property worth	.179	.097	.248	.129	.195	.098	.171	.087	.150	.081
Assessed value	.340	.212	.232	.129	.187	.095	.389	.221	.151	.0711
Bluff distance	.161	.095	.114	.055	.252	.143	.216	.123	.138	.059
Beach lost	.222	.103	.114	.052	.138	.065	.215	.131	.230	.101
Total damage	.183	.109	.213	.125	.096	.049	.187	.081	.111	.057
Total cost	.100	.059	.172	.108	.183	.089	.187	.096	.157	.071

We find Method 2 to be better than Method 1; Method 3, an average of Methods 1 and 2, is found to be better than either Method 1 or 2.

11.3 METHOD 3

Averaging the estimators proposed in Methods 1 and 2 results in the best of the 4 methods of estimation considered. We approximated the population c.d.f. of reach 1 with the following c.d.f.:

$$1/2\ \Phi_{\bar{x}_1, s_1}(\cdot) + 1/2\ \Phi_{\bar{x}, s}(\cdot)$$

where $\Phi_{\bar{x}_1, s_1}(\cdot)$ is the c.d.f. of a Normal $(\bar{x}_1,\ s_1^2)$ variable, etc.

Such a c.d.f. estimator results in a frequency estimator for (a, b) which is the average of Methods 1 and 2:

$$1/2\left[\Phi\left(\frac{\log b - \bar{x}_1}{s_1}\right) - \Phi\left(\frac{\log a - \bar{x}_1}{s_1}\right) + \Phi\left(\frac{\log b - \bar{x}}{s}\right) - \Phi\left(\frac{\log a - \bar{x}}{s}\right)\right].$$

We evaluate the method's accuracy with a simultaneous plot of the following for each of the five stratified random samples:

$$\frac{i - 1/2}{N} \quad \text{vs.} \quad 1/2\ \Phi\left(\frac{x_{(i)} - \bar{x}_1}{s_1}\right) + 1/2\ \Phi\left(\frac{x_{(i)} - \bar{x}}{s}\right) \qquad i = 1, \ldots, N.$$

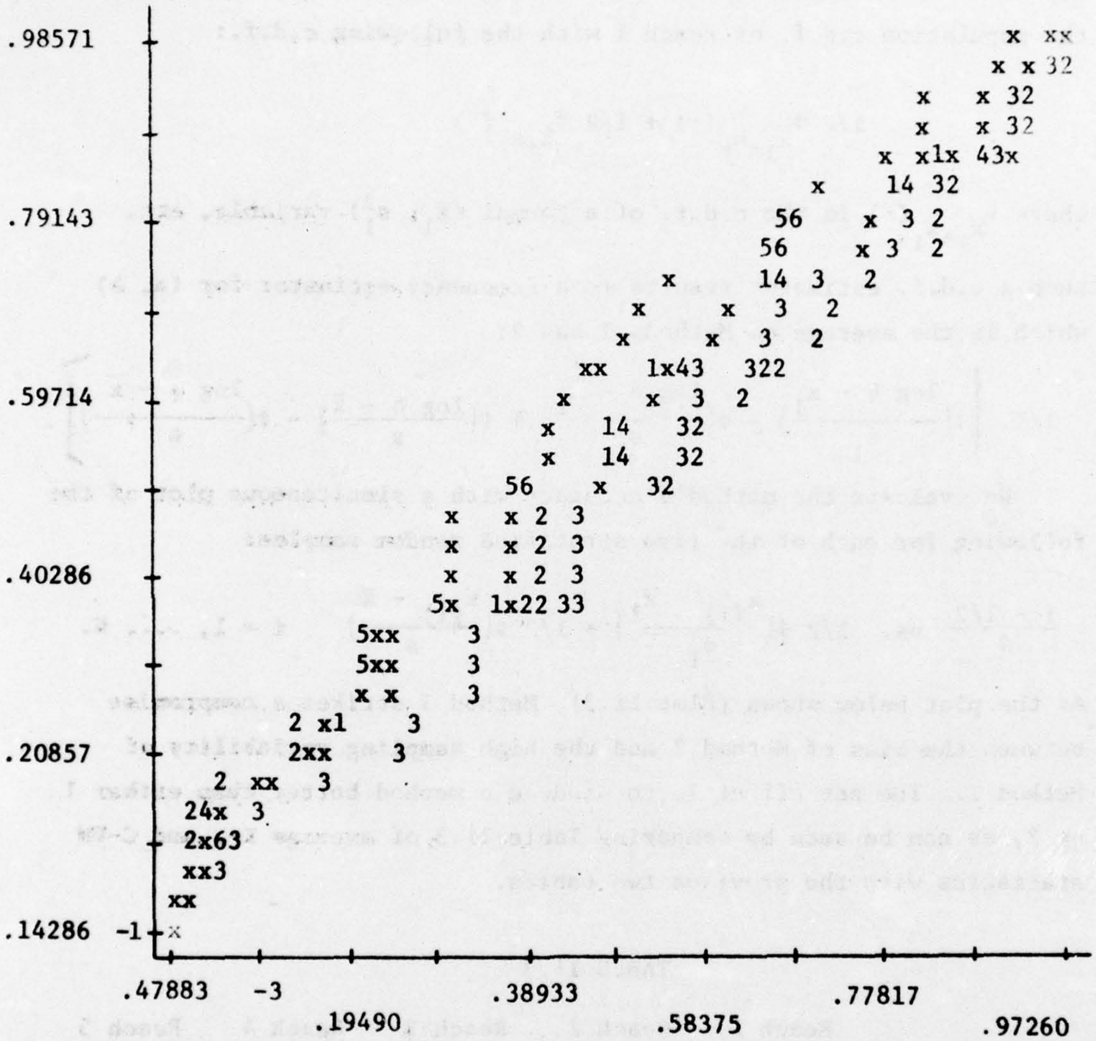
As the plot below shows (Plot 11.3), Method 3 strikes a compromise between the bias of Method 2 and the high sampling variability of Method 1. The net effect is to produce a method better than either 1 or 2, as can be seen by comparing Table 11.3 of average K-S and C-VM statistics with the previous two tables.

TABLE 11.3

	Reach 1		Reach 2		Reach 3		Reach 4		Reach 5	
Variable	K-S	C-VM	K-S	C-VM	K-S	C-VM	K-S	C-VM	K-S	C-VM
Property worth	.139	.080	.184	.083	.168	.086	.185	.089	.106	.055
Assessed value	.172	.122	.168	.088	.186	.085	.353	.178	.140	.066
Bluff distance	.141	.071	.120	.057	.198	.108	.167	.097	.161	.090
Beach lost	.237	.113	.114	.047	.123	.060	.140	.070	.149	.065
Total damage	.141	.079	.123	.068	.124	.064	.159	.072	.120	.059
Total cost	.142	.090	.092	.049	.189	.0948	.128	.074	.165	.071

SCATTER PLOT STRATUM = 1

VAR 6 35 CASES FOR THIS GRAPH



(1) VAR 36 (2) LTOTDAM (3) LTOTDAM (4) LTOTDAM

(5) LTOTDAM (6) LTOTDAM

METHOD #3

PLOT 11.3

The optimal weighting of Methods 1 and 2 has not been determined; however, the following c.d.f. did not perform nearly as well as Method 3:

$$1/5\ \Phi_{\bar{x}_1, s_1}(\cdot) + 4/5\ \Phi_{\bar{x}, s}(\cdot)\ .$$

11.4 METHOD 4

This method makes use of an average of stratum 1 sample data and overall county sample data to estimate the parameters of the fitted normal distribution. Stratum 1 population c.d.f. is estimated by the c.d.f. of a Normal ($\hat{\mu} = 1/2\ \bar{x}_1 + 1/2\ \bar{x}$, $\hat{\sigma}^2 = 1/2\ s_1^2 + 1/2\ s^2$) random variable. Such a c.d.f. estimator results in a frequency estimator for (a, b) of the form:

$$\Phi\left(\frac{\log b - \hat{\mu}}{\hat{\sigma}}\right) - \Phi\left(\frac{\log a - \hat{\mu}}{\hat{\sigma}}\right)$$

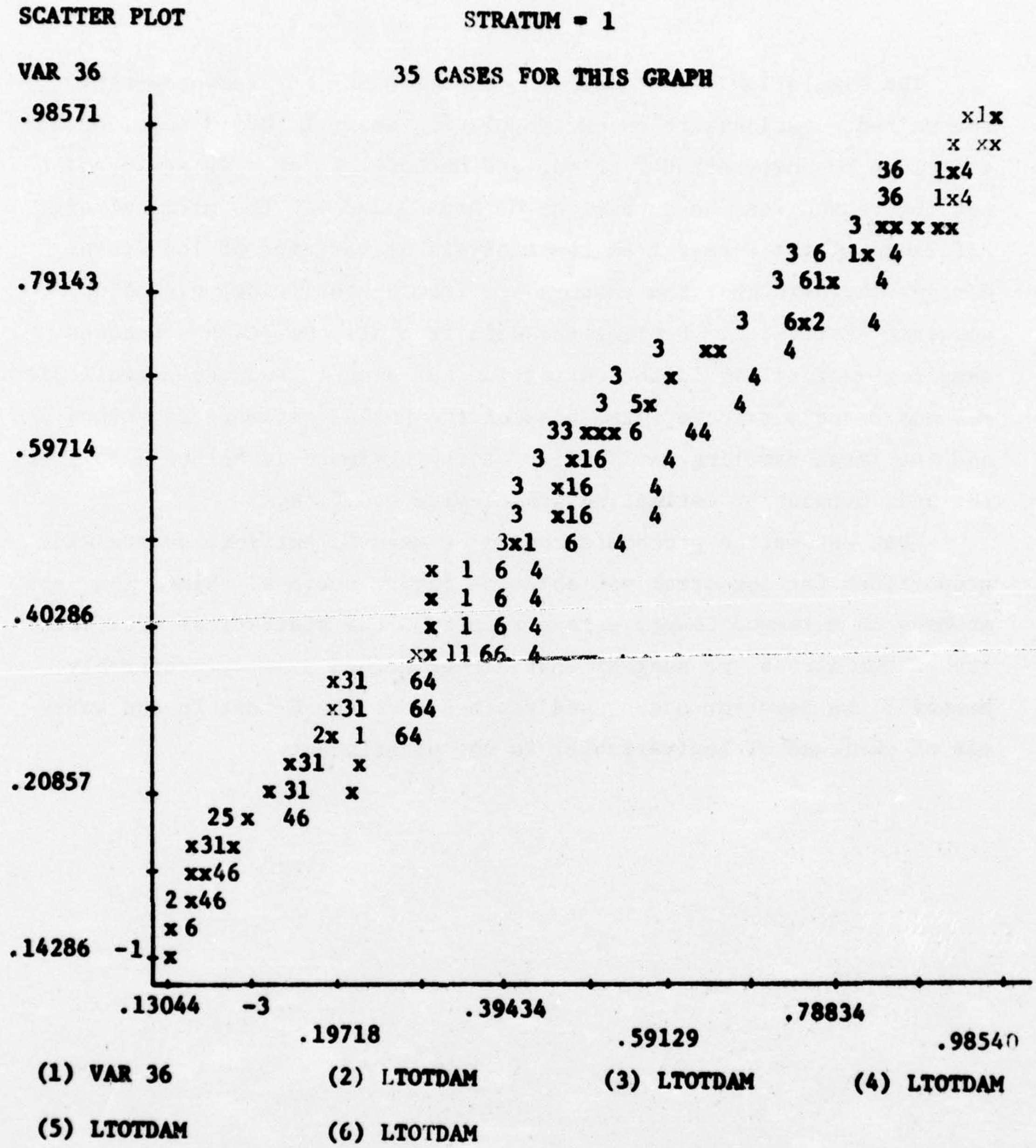
We determine the method's accuracy by plotting the following (Plot 11.4) for each of the five stratified random samples:

$$\frac{i - 1/2}{N} \quad \text{vs.} \quad \Phi\left(\frac{x_{(i)} - \hat{\mu}}{\hat{\sigma}}\right) \qquad i = 1, \ldots, N.$$

Although this method performed better than the first two, it did not fit quite well as Method 3, as we see from the average K-S and C-VM statistics.

TABLE 11.4

	Reach 1		Reach 2		Reach 3		Reach 4		Reach 5	
Variable	K-S	C-VM	K-S	C-VM	K-S	C-VM	K-S	C-VM	K-S	C-VM
Property worth	.159	.085	.153	.063	.185	.115	.167	.085	.125	.065
Assessed value	.190	.118	.145	.084	.207	.100	.356	.169	.142	.071
Bluff distance	.165	.100	.123	.063	.192	.110	.182	.113	.135	.069
Beach lost	.184	.084	.125	.059	.146	.077	.298	.177	.133	.066
Total damage	.136	.078	.163	.088	.111	.053	.217	.101	.132	.066
Total cost	.086	.047	.126	.066	.180	.080	.201	.110	.171	.079



METHOD #4

PLOT 11.4

The simulation of our sampling scheme using the respondents to the mailed questionnaire as our population shows Method 3 best, Method 4 a close second, Method 2 third, and Method 1 worst. We again point out the reason for the success of Methods 3 and 4. The marginal significance of the F-test from the analysis of variance of log (total damage) suggests that the reaches are from a statistical viewpoint somewhat artificial. Pooling the data from all the reaches reduces sampling variability in the estimation but also introduces a small bias. Methods 3 and 4 trade-off the bias of the pooled estimate in Method 2 and the large sampling variability of the estimate in Method 1 to give the best methods of estimating strata-wide c.d.f.'s.

What estimation procedure should be used to estimate strata-wide proportions for lognormal variables of future surveys? Since the best methods in Muskegon County were a result of the statistical artificiality of the strata, we suggest that either Method 3 or 4 (preferably Method 3) be used for uncensused reaches when the F-test in the analysis of variance of log(variable) is not significant.

12.0 PREDICTING FUTURE WATER LEVELS

A time series is a sequence of measurements indexed by time. Our example is the series of 115 average yearly lake levels at Harbor Beach, Michigan from 1860 to 1974, denoted as $\{Y_t: t = 1860, \ldots, 1974\}$. For each of the next 5 years we would like to estimate the marginal probability that the average yearly water level exceeds a certain critical level. More specifically, if

A_i = {event that the average lake level for year i exceeds a certain critical level}

for i = 1975, ..., 1979 then we would like to estimate Prob $\{A_i\}$ for each i. Note that Prob $\{A_i\}$ is a marginal probability since each event A_i refers to its intended year i only and makes no mention of whether the critical level was or was not exceeded for any other year. These probabilities are to be used in a cost benefit analysis.

In order to estimate these probabilities, a model is necessary. A spectral analysis* performed on the residuals of the time series, i.e. $\{R_t = Y_t - \overline{Y} \quad t = 1860, \ldots, 1974\}$, indicated that the following model might be used to "explain" the residuals:

$$R_t = 1.2069 * R_{t-1} - .35264 * R_{t-2} + \varepsilon_t$$

where ε_t is an error term. Such a model is called second order autoregressive model becuase R_t is regressed (least squares) on the two preceding values of the time series. The model above is also stationary, meaning that this model is characteristic of a residual series that has reached some form of steady state or equilibrium in the sense that statistical properties of the series are independent of time. The fitted model above has an R^2 value of .827 (82.7% of the total sums of squares of the residuals is explained by the model). Higher order autoregressive models do not increase the R-squared statistic appreciably. Such a model is fit using the following Midas commands where V1 is the time series $\{Y_t\}$:

---

* Jenkins and Watts, Spectral Analysis and Its Applications, Holden-Day (1969), pages 258-309.

```
TRANS V2=ZEROMEAN (V1)  LABEL=RESID
REGRESSION V=2; 2(-1), 2(-2)  OPTION=MEANZERO
```

A plot of the estimated errors in the model above, i.e.

$$\{\hat{\varepsilon}_t = R_t - 1.2069 * R_{t-1} + .35264 * R_{t-2} \quad : \quad t = 1862, \ldots, 1974\},$$

suggests assuming a normal error structure for the model. Such a normality assumption is the basis for our estimates of the probability of exceeding some fixed level.

Estimation of the probability that the 1975 lake level exceeds some value "a" is computed as follows. Suppose $\hat{R}_t(1) = 1.2069 * R_{t-1} - .35264 * R_{t-2}$ (prediction one year in advance). Then, our opinion about the 1975 lake level can be expressed by a Normal ($\hat{\mu}_{75}$, $\hat{\sigma}^2_{75}$) distribution where

$$\hat{\mu}_{75} = \hat{R}_{1975}(1) + \overline{Y}$$

$$= 1.2069 * R_{1974} - .35264 * R_{1973} + \overline{Y}$$

$$\text{and} \quad \hat{\sigma}^2_{75} = 1/113 \sum_{t=1862}^{1974} [R_t - \hat{R}_t(1)]^2 = 1/113 \sum_{t=1862}^{1974} \hat{\varepsilon}^2_t$$

The estimate $\hat{\mu}_{75}$ is the least squares predicted residual plus $\overline{Y}$ and $\hat{\sigma}^2_{75}$ is the variance in predicting one year in advance. Theoretical justification ("Bayesian predictive distributions") can be given for expressing an opinion about a future observation (average 1975 lake level) with a probability distribution. Rather than doing this, however, we give an intuitive justification for such an expression of opinion. Consider the errors in predictions one year in advance written as

$$Y_t - [\hat{R}_t(1) + \overline{Y}] = Y_t - (1.2069 * R_{t-1} - .35264 * R_{t-2} + \overline{Y})$$

$$= R_t - \hat{R}_t(1) = \hat{\varepsilon}_t$$

for t = 1862, ..., 1974. A histogram of $\{\hat{\varepsilon}_t\}$ appears normally distributed. This suggests that our error in predicting Y_{1975} (this error is a random variable since the average 1975 lake level has not yet

been realized) may also be assumed to a normal random variable with mean and variance given by the mean and variance of $\{\varepsilon_t\}$. However, the mean and variance of $\{\hat{\varepsilon}_t\}$, the errors from least squares regression, are 0 and $\hat{\sigma}^2_{75}$ respectively. If Y_{1975} is the average lake level for 1975 then it is reasonable to assume that $Y_{1975} - \hat{\mu}_{75}$ has a Normal $(0, \hat{\sigma}^2_{75})$ distribution or that Y_{1975} represents a random observation from a Normal $(\hat{\mu}_{75}, \hat{\sigma}^2_{75})$ distribution. Making use of this assumption, the probability that the 1975 lake level exceeds "a" or Prob$\{Y_{1975} > a\}$ is estimated by the probability a Normal $(\hat{\mu}_{75}, \hat{\sigma}^2_{75})$ random variable exceeds "a." After standardization this is written as

$$1 - \Phi\left(\frac{a - \hat{\mu}_{75}}{\hat{\sigma}_{75}}\right).$$

Estimation of the probability that the 1976 lake level exceeds "a" (saying nothing about what happened in 1975 so that a marginal probability is being estimated) is as follows: let

$$\hat{R}_t(2) = 1.2069 * \hat{R}_{t-1}(1) - .35264 * R_{t-2}$$

where R_{t-1} is replaced by a 1-year-in-advance predicted value so that $\hat{R}_t(2)$ predicts R_t based on data $R_{t-2}, R_{t-3}, \ldots$, which is two or more years in advance. A histogram of $\{R_t - R_t(2): t = 1863, \ldots, 1974\}$ appears normally distributed and justifies the approximate normality assumption for the errors in such two year predictions. Because of such a histogram, our opinion about the 1976 lake level can be expressed by a Normal $(\hat{\mu}_{76}, \hat{\sigma}^2_{76})$ distribution where $\hat{\mu}_{76} = \hat{R}_{1976}(2) + \overline{Y}$ and $\hat{\sigma}^2_{76} = 1/112 \cdot \sum_{t=1863}^{1974} (R_t - \hat{R}_t(2))^2$. The estimated probability of exceeding water level "a" in 1976 is $1 - \Phi\left(\frac{a - \hat{\mu}_{76}}{\hat{\sigma}_{76}}\right)$.

To predict the 1977 level we predict three years in advance as follows: $\hat{R}_t(3) = 1.2069 * \hat{R}_{t-1}(2) - .35264\, \hat{R}_{t-2}(1)$ predicts R_t three years in advance. The normal appearance of the histogram of $\{R_t - \hat{R}_t(3): \quad t = 1864, \ldots, 1974\}$ suggests using a Normal $(\hat{\mu}_{77}, \hat{\sigma}^2_{77})$

distribution where $\mu_{77} = \hat{R}_{1977}(3) + \overline{Y}$ and

$$\hat{\sigma}^2_{77} = 1/111 \sum_{t=1864}^{1974} \left(R_t - \hat{R}_t(3)\right)^2$$

to approximate our opinion about the 1977 water level.

Such normally distributed opinions can be computed an arbitrary number of years in advance with increasing variance estimates for their distributions, i.e., $\hat{\sigma}^2_{75} \le \hat{\sigma}^2_{76} \le \hat{\sigma}^2_{77} \le \ldots$. We emphasize, however, that the reliability of estimates computed from such normally distributed opinions is directly related to the number of years ahead we are predicting. Regretfully, prediction more than five years in advance using this method is unreliable.

We again point out the marginal nature of the estimates, i.e., each estimate refers only to the probability of exceeding for its intended year.

Except for the variance estimates, all computations involved here can be easily done on a calculator. The following sequence of Midas commands will compute $\hat{\sigma}^2_{75}, \hat{\sigma}^2_{76}, \ldots, \hat{\sigma}^2_{84}$:

```
TRANS V3=LINEAR VARIABLE=2(-1), 2(-2) RELATION=1.2069, -.35264 L=PRED1YR
TRANS V4=LINEAR V=3(-1), 2(-2) REL=1.2069, -.35264 L=PRED2YR
TRANS V5=LINEAR V=4(-1), 3(-2) REL=1.2069, -.35264 L=PRED3YR
          .                             .
          .                             .
          .                             .
          .                             .
TRANS V12=LINEAR V=11(-1), 10(-2) REL=1.2069, -.35264 L=PRED10YR
TRANS RESULT=13-22 FUNCTION=SUBTRACT VARIABLE=3-12;2 L=*
COMPUTE @A RES=1-10 FUN=POWER(2) V=13-22 L=* STRATA=NONE
SET @A
WRITE * V=1-10
```

The one case written for variable #1 is $\hat{\sigma}^2_{75}$, for variable #2 is $\hat{\sigma}^2_{76}$, etc.

13.0 SUMMARY AND CONCLUSIONS

We summarize the main points of our recommendations below.

Prelist

Adequate care should be taken in preparation of the prelist. A poor prelist may seriously bias estimates of total losses suffered within a county. See the introduction and section 8 for relevant discussion.

Questionnaires and Coding, Section 2

1. Questions asked on the mailed questionnaire should be asked in identical form on the interview to make comparisons of the two questionnaires more meaningful.
2. Questions should be revised so that the following three responses are distinguishable: missing datum, value of zero, inability to answer the question.
3. Various questions should be eliminated, modified, or added to improve the quality of information obtained from questionnaires.
4. Possible responses to a categorical question should be listed on the questionnaire in a mutually exclusive and exhaustive fashion to remove difficulties in interpreting respondents' answers.

Distribution of Variables From Mailed Questionnaire for Six Counties, Section 3

1. Fourteen variables have a multinomial distribution.
2. The non-zero and non-missing values of the following ten variables can be described reasonably well with a lognormal distribution: assessed value, property depth, property worth, total damage, total cost, bluff height, beach depth, bluff lost, bluff distance, and beach lost.

Use of the Lognormal Approximation, Section 4

1. The sample mean and variance of a variable which is approximately lognormal provide misleading information regarding shape and skewness of the variable's distribution. However, the mean and variance of the log of the variable describe the variable's distribution completely.
2. Since the log of a variable which is approximately lognormal is approximately normal, the log of a lognormal type variable may be appropriately used in all statistical procedures based on the assumption of normality. Such procedures used during the course of this analysis include: regression analysis, paired and two-sample t-tests, and analysis of variance.
3. Lognormal models can be used to construct tolerance intervals and estimate population proportions within specified ranges.
4. Lognormal fits within individual reaches for the ten lognormal type variables were good when reaches were large enough. For small reaches we suggest in our sampling scheme that a census be taken so that fitting a distribution is not necessary.

Outliers, Section 5

All outliers should be carefull checked for coding errors, key-punching or typing errors, and response errors. A call or visit to respondent may be necessary to check the validity of a response if an outlier is found not to be the result of a clerical error.

Mailed Questionnaire vs. Personal Interview, Section 6

1. **So that all differences in responses between mailed questionnaire and personal interview will be attributable to differences in techniques of obtaining information, wording of questions should be identical in both settings.**
2. **No statistically significant differences between responses to mailed questionnaire and personal interview were found for the variables tested.**

3. When measures of central tendency for questionnaire and interview data were compared, respondents were found generally to give higher answers on the mailed questionnaire than during the interview for all four damage and loss variables: total damage, total cost, bluff loss, and beach loss.
4. Those who will use these results to make decisions must decide whether responses to the mailed questionnaire give an accurate picture of actual damages suffered or if these responses might tend to be somewhat higher than actual damages suffered from high lake levels. The presence of an interviewer seems to have a conservative effect on answers to many questions.

Respondents vs. Nonrespondents, Section 7

1. No statistically significant differences between answers given by respondents and nonrespondents in the personal interview setting could be found for the variables tested.
2. When measures of central tendency for respondents and nonrespondents were compared, respondents were found generally to give larger answers for property worth, property depth, beach depth, and total cost than did nonrespondents. Nonrespondents tended to report greater beach and bluff losses. Thus, these results indicate that nonrespondents tend to have smaller, less valuable properties than respondents, and to have suffered greater beach and bluff losses from the high lake levels.

Estimates of Totals for Six Counties, Section 8

Estimates based on responses to the mailed questionnaire of actual damages suffered (total damage, total cost, bluff loss, and beach loss) were made for Alcona, Chippewa, Huron, Manistee, Muskegon, and Schoolcraft Counties.

Sampling Plan, Section 9

1. We suggest two possible sampling plans:

   I. Stratified (by reach) simple random sampling for each county.

   II. Stratified (by reach) systematic sampling if we would like, for instance, for the sample to be evenly distributed along the shoreline.

2. Sample size within a particular reach is computed as

   s = maximum (20% of reach size, 30)

   sample size = minimum (x, reach size)

3. Care must be taken to see that there are no nonrespondents in a random sample so that meaningful analyses can be made. Nonrespondents should not be ignored or replaced in the sample--otherwise the randomness of the sample is destroyed and interpretations of results are confused.

4. A census of outliers not found to be coding errors should be conducted.

5. Personal interviews should be given to 40 randomly chosen respondents to the mailed questionnaire, and comparisons such as in section 6 made.

6. In view of the present high water levels in the Great Lakes, we propose that our sampling plan be used on as many (preferably all) remaining counties bordering on the Great Lakes as possible. Cross-county comparisons would not be possible otherwise. Since response rates to the questionnaires may drop appreciably with water levels if the sampling process were to continue over a period of years, we recommend that all sampling be completed as soon as possible.

Estimation of Totals for Future Surveys, Section 10

Eight methods of estimating strata-wide totals for total damage total cost, bluff lost, and beach lost based on a 20% stratified random sample were considered. A procedure for choosing "best" methods of estimating total losses for future surveys is

described.

Estimation of Proportions for Future Surveys, Section 11

Four methods for estimating strata-wide proportions were considered for the approximately lognormal variables property worth, assessed value, bluff distance, beach loss, total damage, and total cost. For each variable the best method involved averaging a reach-wide lognormal fit and a county-wide lognormal fit and using this average distribution to estimate the distribution of the reach population.

Predicting Future Water Levels, Section 12

The following second order stationary autoregressive model was used to describe the residuals

$\{R_t = Y_t - \overline{Y} \quad : \quad t = 1860, \ldots, 1974\}$

of the time series $\{Y_t\}$, the average yearly lake levels at Harbor Beach, Michigan:

$$R_t = 1.2069 * R_{t-1} - .35264 * R_{t-2} + \varepsilon_t.$$

Using the normality of the error structures from this model, our opinions about the average yearly lake levels for subsequent years may be described individually using normal distributions. Estimates of marginal probabilities of exceeding a fixed value for subsequent years are computed from such normal distributions.

APPENDIX V-a

Class*	Value	Variable Name	Class	Value	Variable Name
A	V1	IDNO1	A	V30	FLOOD-H
A	V2	ASSESVAL	A	V31	FLOOD-J1
A	V3	ASSESDAT	A	V32	FLOOD-J2
A	V4	RATIO	A	V33	ERODE-A
A	V5	FRONTAGE	A	V34	ERODE-B
A	V6	PROPDEPT	A	V35	ERODE-C
A	V7	REACHNO	A	V36	ERODE-D
C	V8	DWELLING	A	V37	IDNO3
C	V9	A6A1	A	V38	ERODE-E
C	V10	A6A2	A	V39	ERODE-F
C	V11	A6A3	A	V40	ERODE-G
A	V12	NODWELL1	A	V41	ERODE-H
A	V13	NODWELL2	A	V42	ERODE-J1
A	V14	NODWELL3	A	V43	ERODE-J2
C	V15	DWELTYP1	C	V44	ACTION-1
C	V16	DWELTYP2	A	V45	DATEACT1
C	V17	DWELTYP3	A	V46	COSTMAT1
A	V18	PROPWORT	A	V47	COSTLAB1
A	V19	PROPDELT	C	V48	SUCCESS1
C	V20	PROTACT?	C	V49	ACTION
A	V21	IDNO2	A	V50	DATEACT2
C	V22	DAMAGE?	A	V51	COSTMAT2
A	V23	FLOOD-A	A	V52	COSTLAB2
A	V24	FLOOD-B	C	V53	SUCCESS2
A	V25	FLOOD-C	C	V54	ACTION-3
A	V26	FLOOD-D	A	V55	DATEACT3
A	V27	FLOOD-E	A	V56	IDNO4
A	V28	FLOOD-F	A	V57	COSTMAT3
A	V29	FLOOD-G	A	V58	COSTLAB3

* Class code C = categorical; A = analytical.

Class	Value	Variable Name
C	V59	SUCCESS3
C	V60	B5
A	V61	NOFLOOD
A	V62	FFMAXHT
C	V63	SEEPAGE?
A	V64	BASEMENT
A	V65	MAXWAVE
A	V66	SHORDIST
A	V67	BEACHLOS
A	V68	BANKLOST
A	V69	BANKHGT
A	V70	BLUFFHGT
A	V71	BEACHDEP
A	V72	BLUFFLOS
A	V73	BLUFFDIS
A	V74	BEACLOST
C	V75	FLOODINS
A	V76	COMPDATE
C	V77	Q-A-FFIP
C	V78	HELPYOUR
C	V79	RESULTS

APPENDIX V-b

Variable Code for Erosion Interviews

	1	- ID1
C	2	- Respondent?
C	3	- How interviewed
	4	- Date of interview
	5	- Q8
	6	- Q9
	7	- Building footage
	8	- Q12
	9	- % of living area occupied
C	10	- Q14
C	11	- Q15
	12	- Property worth
C	13	- Method of estimation
	14	- Date of estimation
	15	- Q17
C	16	- Method of estimation
	17	- Date of estimation
	18	- Q19a - structures
	19	- Q19a - contents
	20	- Q20a
	21	- ID2
	22	- Q19b - structure
	23	- Q19b - contents
	24	- Q20b
	25	- Q19c - stairs
	26	- Q20c
	27	- Q19d - lawns
	28	- Q20d
	29	- Q19e - other
	30	- Q20e
C	31	- Q21
	32	- Income lost
C	33	- Action #1
	34	- Date of action
	35	- Cost of action
C	36	- Success of action
	37	- Action #2
	38	- Date of action
	39	- ID#3
	40	- Cost of action
C	41	- Success of action
	42	- Q24 - depth
	43	- Q25 - stairs
	44	- Q25 - cost to replace
	45	- Q26 - septic

46 - Q26 - cost to replace
47 - Q27 - residence
48 - Q27 - house
49 - Q27 - contents
C 50 - Q28 - Building type #1
51 - Q28 - feet
52 - Q28 - cost
C 53 - Q28 - Building type #2
54 - Q28 - feet
55 - Q28 - cost
C 56 - ID#4
C 57 - Q28 - Building type #3
58 - Q28 - feet
59 - Q28 - cost
60 - Q29 - 25 bluff
61 - Q30 - 50 bluff
62 - Q32 - 100 bluff
63 - Q33 - 125 bluff
64 - Q34 - 150 bluff
65 - Q35 - 200 bluff
66 - Q36 - other
67 - Q37 - bluff height
68 - Q38 - beach loss
69 - Q39 - beach depth
70 - Q40 - bluff loss
71 - Q41 - bluff to foundation
72 - Q42 - total frontage
73 - ID#5
74 - Q43 - beach length
75 - Q44 - bluff length
C 76 - Residence type
C 77 - Flood insurance
78 - How long owned/resided

"C" denotes categorical variables

Variable Code for Flood Interviews

1 - ID1
C 2 - Respondent?
C 3 - How interviewed
4 - Date of interview
5 - Q8A
6 - Q8B
7 - Q8C
8 - Q9 - footage
9 - Q11
10 - Q12 - %
11 - Q13 - years occupied
12 - Q14a - No. of flood events
13 - Q14b - date of flood
14 - Q14b - date of flood
C 15 - Q15 - property use
C 16 - Q16 - ownership
17 - Q17a - Property worth
C 18 - Method of estimation
C 19 - Current estimation?
20 - Year of estimation
21 - Q17b - property worth, low levels
22 - ID2
23 - Q18a - developed property worth
C 24 - Method of estimation
C 25 - Current estimation?
26 - Year of estimation
27 - Q18b - developed property worth, low levels
28 - Q20 - flood date
29 - Q21 - hours of warning
30 - Q23a
31 - Q23b
32 - Q23c
33 - Q23d
34 - Q25b - structure
35 - Q25b - contents
36 - Q25b - structure, others
37 - Q25b - contents, others
38 - Q25b - vehicles
39 - Q25b - lawns
40 - ID3
41 - Q25b - clean-up
42 - Q25b - other
43 - Q26
44 - Q27 - net income lost
C 45 - Action #1

46 - Date of action 1
47 - Cost of action 1
C 48 - Success of action 1
C 49 - Action #2
50 - Date of action 2
51 - Cost of action 2
C 52 - Success of action 2
C 53 - Action #3
54 - Date of action 3
55 - Cost of action 3
C 56 - Success of action 3
57 - Q30 - height of storm water
58 - Q30a - date
59 - Q31 - depth of basement
C 60 - Q31a - seepage
61 - ID4
62 - Q32 - wave height
63 - Q33 - distance to shore
64 - Q34 - beach lost
65 - Q35 - bank lost
66 - Q36 - bank height
67 - Q37 - total length of shoreline
68 - Q38 - beach length
69 - Q39 - bank length
70 - Q40 - property width
C 71 - Flood insurance?
C 72 - Residence type

"C" denotes categorical variables

APPENDIX V-c

Great Lakes Shoreline Damage Questionnaire

This questionnaire is designed to find out the amount of damage suffered by lakeshore property owners from high water levels, the attempts property owners have made to protect their homes and land, and the effectiveness of these actions. Many questions may be answered simply by marking an "X" in the appropriate box. An arrow leading from an answer box indicates the next question to be answered. If there is no arrow or skip instruction, please go on to the next question.

LOCATION

A1. We need to be able to study the effects of high water and marine weather conditions on <u>specific areas</u>. Your answers to the first series of questions will help us locate your property. What is your <u>lakeshore</u> mailing address?

NAME: ______________________________

ADDRESS: ______________________________

____
(CITY) (COUNTY) (STATE) (ZIP)

A2. Does your property front on the lake?

☐ YES ☐ NO

YES	NO
A2a. What is the length of your shoreline frontage? ________ FEET	A2c. About how far from the lake is your property line at the nearest point? ________ FEET
A2b. How many feet back from the present shoreline does your property extend? (Approximate "depth") ________ FEET	A2d. What is the approximate size of your property, that is, how many feet does it measure each way? ________ FEET BY ________ FEET

A3. Please indicate the location of your property as accurately as you can <u>on the enclosed map</u>, by making an "X" at the spot where your property is.

A4. What is the name of the public road nearest your property?

____

This road is to the *(North/South/East/West)* of your property, and:
(CIRCLE ONE)

(CHECK ONE)

either ☐ touches your property

or ☐ is ___________ feet from your property

A5. What is the name of the road which intersects the above road nearest your property?

____

This road is to the *(North/South/East/West)* of your property, and:
(CIRCLE ONE)

(CHECK ONE)

either ☐ touches your property

or ☐ is ___________ feet from your property

A6. Do you have a house cottage or any other dwelling units on your pro

☐ **YES** → Continue with A6a on the next page.

☐ **NO** → Skip to A7 on the next page.

LIST EACH DWELLING STRUCTURE IN A SEPARATE COLUMN

A6a. What kind of dwelling(s) do you have? (For example, house, cottage, mobile home, apartment house, etc.)	______	______	______
A6b. How many dwelling units are in each structure? (single family, duplex, 6 apartments, etc.)	______	______	______
A6c. How is the structure used? (CHECK AS MANY AS APPLY FOR EACH STRUCTURE)			
AS A SEASONAL RESIDENCE	☐	☐	☐
AS A PERMANENT RESIDENCE	☐	☐	☐
AS INCOME PROPERTY	☐	☐	☐

A7. These next questions are to determine what effects, if any, changes in lake levels have on property value. If you were to sell your property now, during high lake levels, how much do you think you could get?

$ ______________

A8. If the lake level were <u>not</u> so high could you sell your property for more, less, or about the same?

☐ **MORE WITHOUT HIGH LAKE LEVELS** ☐ **ABOUT THE SAME** ☐ **LESS WITHOUT HIGH LAKE LEVELS**

A8a. How much more?

$ ______________

A8b. How much less?

$ ______________

A9. Have you taken any protective action because of recent high water levels, or have you suffered any damage, or is there risk of damage to your lakeshore property?

☐ **YES** → **Continue to next page**

☐ **NO** → **Turn to page 8**

DAMAGE ESTIMATE

In order to estimate the damage caused by high waters (either through flooding or erosion) we need to get your best estimate of all costs you have incurred from Labor Day, 1972 until Labor Day, 1974.

B1. Has your shoreline property sustained any actual damage due to high lake levels since Labor Day, 1972?

☐ YES ☐ NO → Skip to B4, Next page

B2. Please indicate the type of damage and the approximate amount of loss in dollars. Enter the amount in the appropriate column according to the type of damage you have suffered.

	DAMAGE CAUSED BY: FLOODING DUE TO HIGH LAKE LEVELS	EROSION OF SHORE LINE
a. Structure and contents of residence	$ ______	$ ______
b. Detatched garages and out buildings	$ ______	$ ______
c. Docks and boathouses	$ ______	$ ______
d. Stairways and ramps	$ ______	$ ______
e. Grounds, landscaping, trees, etc.	$ ______	$ ______
f. Clean up costs	$ ______	$ ______
g. Septic system	$ ______	$ ______
h. Loss of rental income	$ ______	$ ______
j. Other (Please specify)		
______	$ ______	$ ______
______	$ ______	$ ______

B3. TOTAL AMOUNT OF DAMAGE $ ______ $ ______

PROTECTIVE ACTIONS

B4. Has any protective action been taken with regard to your property since Labor Day, 1972? (Typical protective actions are relocating your house or cottage, or building revetments, sea walls, groins, etc.)

☐ YES ☐ NO ——▶ Skip to B5, Bottom of page

If any action was taken by a group of neighbors or an association, please write here the name and address of an officer or leader of that group so that we can contact that person directly. List below only what you did. ________________

B4a. What did you do? Please list each thing in a separate column at the right.			
B4b. Give the approximate date that you did this. (If work took some time show approximate starting and ending months.)	______ (MONTH) (YEAR)	______ (MONTH) (YEAR)	______ (MONTH) (YEAR)
B4c. What did it cost? MATERIALS	$ ______	$ ______	$ ______
LABOR (If you did some of the work, estimate what you would have had to pay to have someone else do it.)	$ ______	$ ______	$ ______
B4d. How successful were these efforts?			

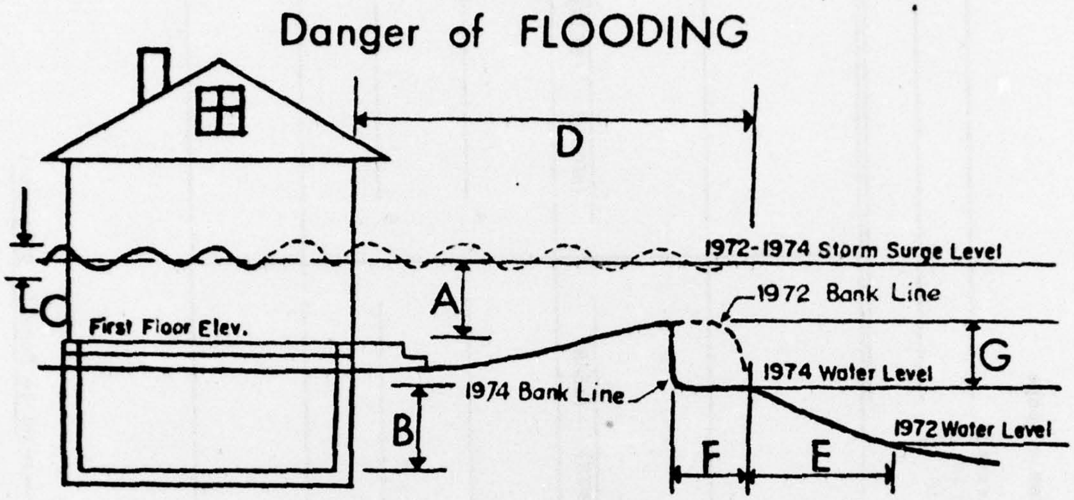
B5. Please check the one which best describes your property.

☐ No dwelling on property ——▶ Go to Page 8

☐ Dwelling(s) in danger of flooding caused by high lake levels ——▶ Go to Page 6, Diagram I

☐ Flooding not a problem, but property threatened by shoreline erosion ——▶ Go to Page 7, Diagram II

DIAGRAM I



In order to get a clear idea of just what flood risk your house is under, we would like you to refer to the above diagram when answering the following questions.

C1. How many times has your residence been flooded since Labor Day, 1972? ____________

C2. Estimate the maximum height above the first floor elevation that the storm driven water has reached on your residence. ____________ FEET distance "A" above

C3. Are you experiencing continuing basement seepage problems?

☐ YES ☐ NO ☐ NO BASEMENT GO ON TO C5

C4. Estimate the number of feet your basement floor is below the existing water level. ____________ FEET distance "B" above

C5. Estimate the maximum height of any wave to date that has acted on your residence. ____________ FEET distance "C" above

C6. Estimate the distance of your residence from the existing shoreline. ____________ FEET distance "D" above

C7. About how many feet of beach have you lost? ____________ FEET distance "E" above

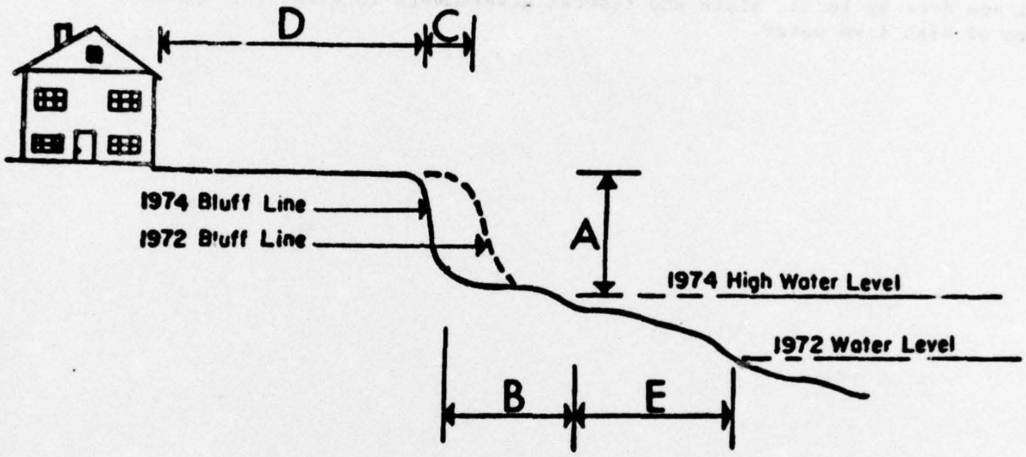
C8. About how many feet of bank have you lost? ____________ FEET distance "F" above

C9. How high is your bank above existing water levels? ____________ FEET distance "G" above

PLEASE TURN TO PAGE 8

DIAGRAM II

Danger of EROSION



In order to get a clear idea of just what risk your house is under, we would like you to refer to the above diagram when answering the following questions.

D1. What is the approximate height of the bluff or embankment above the existing water level? ______ FEET distance "A" above

D2. How deep is your present beach? ______ FEET distance "B" above

D3. What is the depth of bluff loss due to erosion since Labor Day, 1972? ______ FEET distance "C" above

D4. Estimate the distance between the edge of the bluff or embankment to the foundations of your house. ______ FEET distance "D" above

D5. How many feet of beach have you lost? ______ FEET distance "E" above

THE QUESTIONS ON THIS PAGE ARE FOR EVERYONE

E1. Do you currently have flood insurance coverage? ☐ YES ☐ NO

E2. Please give any additional comments below, or tell us what you would like to see done by local, state and federal governments to alleviate the problem of high lake water.

E3. What is the date this form was completed? ______ (MONTH) (DAY) (YEAR)

E4. Please give us your permanent mailing address if it is different from the one given on page one.

☐ SAME AS ADDRESS GIVEN

NAME ______

ADDRESS ______

______ (CITY) (COUNTY) (STATE) (ZIP)

TELEPHONE ______ (AREA CODE) ______

Thank you very much for your cooperation. Please return this questionnaire in the post-paid envelope provided.

To express our appreciation for your help we would like to send you a copy of some of the major findings of this study. If you would like to see these results, just check the appropriate box below. For your convenience we have also listed several other publications which you may request.

☐ Questions and Answers on the Federal Flood Insurance Program

☐ Brochure entitled *Help Yourself* discussing different methods of shore protection

☐ Report on the results of this study

APPENDIX V-d

PLEASE

PRINT

RESPONSES

ID Code No. ____________

Sub Code No. ____________

PERSONAL INTERVIEW FORM

BLUFF EROSION DAMAGES - RESIDENTIAL PROPERTIES

1. Great Lakes ______________________ 2. State ______________

3. County Reach __________ 4. Interviewer ______________ 5. Date __________

6. Property Owner ________________________________ Telephone No. __________

Address ____

____

7. Respondent's Name ________________________________ Telephone No. __________

Address ____

____

The following questions are to determine what effects, if any, changes in lake levels have had on your property. Specifically, information is being gathered for the period Labor Day 1972 to Labor Day 1974.

8. How many housing units occupy this property? ______________
   (Where more than one, complete a separate interview form for each housing unit, where possible.)

9. How many of the housing units just mentioned are subject to risk from erosion? ______________

10. What is the total square footage of the building in which you live at the lakeshore? ____________ sq. ft.

    IF DON'T KNOW GO TO QUESTIONS 11 & 12

    IF FIGURE IS PROVIDED GO TO QUESTION 13

11. What are the dimensions of the building in which you live at the lakeshore, that is

    a. What is the length ____________ ft.

    b. What is the width ____________ ft.

12. How many stories are there in this building? __________

13. How much of the living area is occupied by you and your dependents? ________________________

PERSONAL INTERVIEW FORM

BLUFF EROSION DAMAGES - RESIDENTIAL PROPERTIES (Cont'd)

14. Is your residence occupied on a

(a.) seasonal basis /__/ or (b.) year round basis /__/

15. Is this residence

(a.) owned by you /__/ (b.) rented by you /__/
(this includes a mortgage)

(c.) other /__/ Describe ______________________

____

The following questions are designed to evaluate the influence of high lake levels on the market value of your property.

16. What is the market value of this property given the high lake levels and existing rates of bluff erosion? $ __________

a. What is the basis or method of your estimate?

Recent Sale /__/ Best Guess /__/ Appraiser's Estimate /__/

Recent Sale of Property Similar to that of Yours /__/

b. Is your estimate made in current dollar terms? /__/ Yes /__/ No

If not, what time period is reflected in your estimate? __________

17. What would you estimate the market value of this property to be if normal lake levels were to return? $ __________

a. What is the basis or method of your estimate?

Recent Sale /__/ Best Guess /__/ Appraiser's Estimate /__/

Recent Sale of Property Similar to that of Yours /__/

b. Is your estimate made in current dollar terms? /__/ Yes /__/ No

If not, what time period is reflected in your estimate? __________

18. In your own words, what would cause the change in market value, if any, with a return to normal lake levels?

Response: ______________________________________

____

PERSONAL INTERVIEW FORM

BLUFF EROSION DAMAGES - RESIDENTIAL PROPERTIES (Cont'd)

The following questions will attempt to obtain a detailed estimate of damages experienced from Labor Day 1972 to Labor Day 1974.

19. Will you please estimate the costs of damages to the following items between Labor Day 1972 and Labor Day 1974?

20. On what dates were these damages experienced?

a. House — ______ mo. ______ yr.

Structure $ ______

Contents $ ______

b. Other Buildings — ______ mo. ______ yr.

Structure $ ______

Contents $ ______

c. Stairways and Walks $ ______ — ______ mo. ______ yr.

d. Lawn $ ______ — ______ mo. ______ yr.

e. Other (Specify)

______ — ______ mo. ______ yr.

______ — ______ mo. ______ yr.

21. Have you lost any rental or other income obtained from the use of your property due to the risk of erosion or because of actual erosion damages?

/___/ Yes /___/ No

What is your estimate of the amount of NET INCOME lost? $ ______

22. The following questions will concern any protective actions you may have taken at any time to reduce the risk of damages to your property, due to high lake levels.

a. Have you physically relocated any buildings /___/ Yes /___/ No (go to question b)

PERSONAL INTERVIEW FORM

BLUFF EROSION DAMAGES - RESIDENTIAL PROPERTIES (Cont'd)

22. a. (1) On what date? ____________ mo. yr.

(2) How much did this cost? $____________ (now go to part b.)

b. Have you taken temporary or emergency protective measures? /__/ Yes /__/ No (go to part c.)

(1) On what date? ____________ mo. yr.

(2) How much did this cost? $ ____________ (now go to part c.)

c. Have you attempted to provide permanent structural protection? /__/ Yes /__/ No (go to question 23)

(1) On what date? ____________ mo. yr.

(2) How much did this cost? $ ____________

IF YES TO QUESTIONS 22a, b, or c:

23. What were the protective actions taken to reduce the risk of damages to your property?

a. Please describe: ____________

____

____

____

IF NO TO QUESTIONS 22a, b, and c:

23. Have efforts been taken at any time to prevent erosion damages? /__/ Yes /__/ No (go to question 24)

PERSONAL INTERVIEW FORM

BLUFF EROSION DAMAGES - RESIDENTIAL PROPERTIES (Cont'd)

INTERVIEWER: (If the responses to questions 22 and 23.a. indicate the construction of a shore protection structure, proceed to part b. below. If no shore protection structures, but some other protective effort taken, go to part e. If no protective actions, go on to question 24)

23. b. Can you furnish us a photograph of your structure that we can attach to this questionnaire?

/___/ Yes (go to part e.)   /___/ No (go to part c.)

INTERVIEWER: (Hand out copy of pictured shoreline protection structures.)

c. Which of the pictured shoreline protection structures most closely looks like your structure?

Picture No. ___________ (go to part e.)   /___/ Structure not pictured (go to part d.)

INTERVIEWER: (Pick up copy of pictured shoreline protection structures.)

d. Would you be willing to sketch the type of protective structure constructed at your property?

/___/ Yes (attach sketch to questionnaire and go to part e.)   /___/ No (go to part e.)

e. Were your efforts successful in preventing erosion damages?

/___/ Yes   /___/ No

Please explain: ____

____

____

____

____

PERSONAL INTERVIEW FORM

BLUFF EROSION DAMAGES - RESIDENTIAL PROPERTIES (Cont'd)

24. What is the total depth of this property? ____________feet

The following questions will ask about damages you would experience if bluff erosion were to continue on into the future.

25. How many feet of additional bluff loss from the existing bluff line is required before stairways and walks are damaged?

____________feet

What would the dollar value of such a loss be or the cost of its replacement?

$____________

26. How many feet of additional bluff loss from the existing bluff line is required before the septic system is damaged?

____________feet

What would the dollar value of such a loss be or the cost of replacement?

$____________

27. How many feet of additional bluff loss from the existing bluff line is required before your residence is damaged?

____________feet

a. What would the dollar value of a complete loss to the structure of your residence be? $____________

b. What would the dollar value of damages to the contents of your residence be if you could not evacuate these in time? $____________

28. How many feet of additional bluff loss from the existing bluff line is required before other buildings are damaged?

What would the dollar value of complete losses to such buildings be?

____________ ________feet $____________
type of bldg.

____________ ________feet $____________
type of bldg.

____________ ________feet $____________
type of bldg.

29. If 25 feet of bluff were lost from the <u>existing</u> bluff line what would the dollar value of damages to lawn, trees, ornamental shrubs, and landscaping be?

$____________

30. If instead 50 feet of bluff were lost from the <u>existing</u> bluff line what would the dollar value of damages to lawn, trees, ornamental shrubs, and landscaping be?

$____________

31. If 75 feet of bluff were lost from the <u>existing</u> bluff line what would the dollar value of damages to lawn, trees, ornamental shrubs, and landscaping be?

$____________

PERSONAL INTERVIEW FORM

BLUFF EROSION DAMAGES - RESIDENTIAL PROPERTIES (Cont'd)

32. **If 100 feet of bluff were lost from the existing bluff line what would the dollar value of damages to lawn, trees, ornamental shrubs, and landscaping be?**

$\$$______________

33. **If 125 feet of bluff were lost from the existing bluff line, what would the dollar value of damages to lawn, trees, ornamental shrubs, and landscaping be?**

$\$$______________

34. **If 150 feet of bluff were lost from the existing bluff line, what would the dollar value of damages to lawn, trees, ornamental shrubs, and landscaping be?**

$\$$______________

35. **If 200 feet of bluff were lost from the existing bluff line what would the dollar value of damages to lawn, trees, ornamental shrubs, and landscaping be?**

$\$$______________

36. **Are there any other items that would be damaged by further bluff loss from the existing bluff line that have not already been discussed?**

/__/ **Yes** /__/ **No** **(go to question 37)**

Describe **(INTERVIEWER:** OBTAIN DISTANCE AND \$ VALUES**)**

____

____

____

____

____

PERSONAL INTERVIEW FORM

BLUFF EROSION DAMAGES - RESIDENTIAL PROPERTIES (Cont'd)

INTERVIEWER: (Hand the respondent the card with the bluff profile diagram on it and read the following statement and questions.)

This standard drawing can be used in a number of ways to answer the following questions. Any use that can be made of this drawing in answering the questions should be made. Penciled modifications which more clearly illustrate your situation should be made directly on the drawing.

Question	Diagram Exhibit	Response
37. What is your estimate of the height of the bluff or embankment above the existing level?	A	______feet
38. What is your estimate of the number of feet of beach width you have lost?	E	______feet
39. What is your estimate of the width of your present beach?	B	______feet
40. What is your estimate of the width of bluff loss due to erosion between Labor Day 1972 and Labor Day 1974?	C	______feet
41. What is your estimate of the distance between the edge of the bluff or embankment to the foundation of your residence?	D	______feet

INTERVIEWER: TAKE BACK CARDS

Question	Diagram Exhibit	Response
42. What is your estimate of the total length of shoreline for this property?	Not dia-gramed	______feet
43. What is your estimate of the length of your shoreline for which beach has been lost?	Not dia-gramed	______feet
44. What is your estimate of the length of your shoreline for which bluff area has been lost?	Not dia-gramed	______feet

45. Do you currently have flood insurance coverage: /__/ Yes /__/ No

46. How long have you owned or occupied your lakeshore property? ______

PERSONAL INTERVIEW FORM

BLUFF EROSION DAMAGES - RESIDENTIAL PROPERTIES (Cont'd)

INTERVIEWER'S SHEET

Record your evaluation of respondents answers to the questions in general and give particular attention to questions concerning market values (questions 16 and 17), cost estimates of damages and protective actions, and estimates of distances.